ELSEVIER

EVOLUTIONARY PROTEIN DESIGN

Volume 55

Frances H. Arnold

ADVANCES IN PROTEIN CHEMISTRY

Volume 55

Evolutionary Protein Design

This Page Intentionally Left Blank

ADVANCES IN PROTEIN CHEMISTRY

EDITED BY

FREDERIC M. RICHARDS
Department of Molecular Biophysics
and Biochemistry
Yale University
New Haven, Connecticut

DAVID S. EISENBERG
Department of Chemistry and Biochemistry
University of California, Los Angeles
Los Angeles, California

PETER S. KIM
Department of Biology
Massachusetts Institute of Technology
Whitehead Institute for Biomedical Research
Howard Hughes Medical Institute Research Laboratories
Cambridge, Massachusetts

VOLUME 55

Evolutionary Protein Design

EDITED BY

FRANCES H. ARNOLD
California Institute of Technology
Pasadena, California

ACADEMIC PRESS
San Diego London Boston New York
Sydney Tokyo Toronto

This book is printed on acid-free paper. ♾

Academic Press
A Harcourt Science and Technology Company
525 B Street, Suite 1900, San Diego, California 92101-4495, USA
http://www.academicpress.com

Academic Press
Harcourt Place, 32 Jamestown Road, London NW1 7BY, UK
http://www.academicpress.com
Library of Congress Catalog Card Number: 0065-3233

International Standard Book Number: 0-12-034255-3

PRINTED IN THE UNITED STATES OF AMERICA
00 01 02 03 04 05 5B 9 8 7 6 5 4 3 2 1

CONTENTS

New Functions from Old Scaffolds: How Nature Reengineers Enzymes for New Functions

Patricia C. Babbitt and John A. Gerlt

Evolution of Protein Functon by Domain Swapping

Marc Ostermeier and Stephen J. Benkovic

Rational Evolutionary Design: The Theory of *In Vitro* Protein Evolution

Christopher A. Voigt, Stuart Kauffman, and Zhen-Gang Wang

Temperature Adaptation of Enzymes: Lessons from Laboratory Evolution

Patrick L. Wintrode and Frances H. Arnold

Structural Analysis of Affinity Matured Antibodies and Laboratory-Evolved Enzymes

M. Cecilia Orencia, Michael A. Hanson, and Raymond C. Stevens

Molecular Breeding: The Natural Approach to Protein Design

JON E. NESS, STEPHEN B. DEL CARDAYRÉ, JEREMY MINSHULL,
AND WILLEM P. C. STEMMER

Analysis of Large Libraries of Protein Mutants Using Flow Cytometry

GEORGE GEORGIOU

From Catalytic Asymmetric Synthesis to the Transcriptional Regulation of Genes: *In Vivo* and *In Vitro* Evolution of Proteins

CARLOS F. BARBAS III, CHRISTOPH RADER, DAVID J. SEGAL, BENJAMIN LIST,
AND JAMES M. TURNER

In Vitro Selection and Evolution of Proteins

Andreas Plückthun, Christiane Schaffitzel, Jozef Hanes, and Lutz Jermutus

PREFACE

Evolutionary methods have emerged in the short space of a decade as key tools for protein design and engineering. The robustness and wide applicability of these methods, together with some impressive demonstrations of altering proteins to achieve defined functional goals, have already placed them in a very visible position in biological design. Laboratory evolution is also beginning to provide new tools for understanding the molecular basis of protein function, structure, and evolution. This volume is a compilation of ideas and recommendations from some of the leading practicioners of evolutionary design. Several chapters also present insights derived from natural protein evolution that both inspire and help guide evolutionary engineering and design.

Classical biochemical approaches to studying components and trying to build up to larger and more complex systems—macromolecular assemblies, genetic circuits, metabolic pathways, cells, organisms—eventually fail somewhere along the line. Even single proteins rarely submit to being designed from the bottom up, and most biological systems are overwhelmingly complex when viewed from this perspective. Evolution, however, works on and draws its strength from this very complexity. The ''blind watchmaker''[1] is the colorful name given to the algorithm of mutation and natural selection that has chugged its way to everything from nitrogenase to the chimpanzee. Stymied in their attempts to tame proteins (and other complex systems) ''rationally,'' scientists and engineers are now developing their own laboratory versions of this algorithm, and have excellent results to show for it.

Applied molecular evolution, directed evolution, molecular breeding—the process is known by different names—all use some variation

[1] R. Dawkins, ''The Blind Watchmaker.'' Longmans, London, 1986.

on a mutation and selection strategy to improve or even create new functions in proteins. A laboratory evolution process, of course, differs from the blind watchmaker in that there is a functional goal. In this sense it is ''directed'' and more like the breeding process practiced by mankind for thousands of years to manipulate crops, flowers, and domestic animals.[2] Breeding molecules in the test tube is much more flexible than breeding plants or animals, however. Useful results can be obtained using a variety of different strategies, as is apparent from the diversity of approaches outlined in this volume. It is also much faster, since a generation can be completed in as little as a day, and with large numbers of progeny.

What problems are amenable to evolutionary approaches? Successful laboratory evolution requires that (1) the desired function be physically feasible, (2) the function be evolutionarily feasible (i.e., that there exists a mutational pathway to get from here to there through identifiable variants), and (3) you can make and screen libraries of mutants complex enough to contain beneficial mutations or recombinations. Ultimately, perhaps the most restrictive requirement is the second one, getting there from here. Test tube evolution will not yield answers to problems for which the paths are highly improbable or nonexistent. For example, new catalytic sites that require the simultaneous placement of several residues are difficult to evolve. (They are also difficult to design rationally!) A stepwise path of evolving binding and then catalysis may yield workable, if not ideal, solutions. So may a combined strategy in which computation, to calculate a poor but workable configuration, is followed by evolutionary fine-tuning. Just now emerging, combined computational–evolutionary approaches will dominate the protein design scene in coming decades.

Daniel Dennett's and Richard Dawkins' books ''Darwin's Dangerous Idea''[3] and ''The Blind Watchmaker''[1] should be read by all students of molecular evolution. Both use Jorge Luis Borges' wonderful short story, the Library of Babel,[4] to introduce the concept of sequence space. The Library of Babel is the vastly (''very much more than astronomically'') large collection of *all possible* books. This unimaginably large collection contains the complete story of your life (and death). It also contains a very large number of imposter biographies, some of which differ from the true one by no more than a single date. This library

[2] The analogy to breeding is well developed at the Maxygen web site (*www.maxygen.com*).

[3] D. C. Dennett, ''Darwin's Dangerous Idea. Evolution and the Meanings of Life.'' Simon & Schuster, New York, 1995.

[4] J. L. Borges, The Library of Babel, in ''Labyrinths: Selected Stories and Other Writings.'' New York: New Directions, 1962.

contains all the information of mankind's history and future, but, alas, does so in no particular order that can be discerned by its long-suffering librarians. Dennett's Library of Mendel is the equivalent collection of all possible gene sequences wherein resides the code for all living things. In fact, the code for *all possible* life is also here, interspersed unfortunately with a great deal of gibberish. Nature has figured out how to traverse the functional volumes and explore new knowledge. We, the librarians of Mendel's collection, are challenged to do the same. Exploring this space promises exciting adventures for years to come, and the solutions we encounter offer a remarkable opportunity to unravel the mysteries of the amazing molecular machines that are proteins.

Frances H. Arnold

This Page Intentionally Left Blank

NEW FUNCTIONS FROM OLD SCAFFOLDS: HOW NATURE REENGINEERS ENZYMES FOR NEW FUNCTIONS

By PATRICIA C. BABBITT* and JOHN A. GERLT†

*Departments of Biopharmaceutical Sciences and Pharmaceutical Chemistry, University of California, San Francisco, California 94143, and †Departments of Biochemistry and Chemistry, University of Illinois at Urbana-Champagne, Urbana, Illinois 61801

I. INTRODUCTION

This volume, devoted to ‘‘Evolutionary Approaches to Protein Design,’’ focuses on the recent applications of new methods in high throughput protein engineering. Why, then, begin the book with a discussion of natural evolution when it is already clear that *in vitro* and other laboratory methods not only hasten the engineering process but explore structural solutions that nature never thought of? Several reasons come to mind. First, understanding how nature has reengineered protein structures for new functions will be useful for intelligent design of characterization protocols for reengineered proteins whether their sources are *in vivo* or *in vitro* evolution. Second, it may not always be desirable to generate all possible variants of a structure but rather to generate a smaller subset specifically targeting some element of function or specificity. Refinement of the engineering strategy through rational design

0065-3233/00 $35.00

principles depends heavily on a sophisticated understanding of the structure-function relationships pertinent to the template being used. In other words, knowing how nature evolved a particular scaffold to deliver specific aspects of function is important. Third, choice of an appropriate template for reengineering is not always straightforward. Analysis of how nature reengineers specific scaffolds for new functions can provide insight into the constraints and engineering opportunities already embedded in candidate template structures. As more superfamilies are characterized, we expect such information to be useful for deciding which scaffold would be the best starting template for a particular engineering application.

A primary issue in examining how nature reengineers protein structures for new functions is to ask how often and for what range of functions each template has been used. Although it is clear that we have only characterized a subset of all of the known or predicted folds, those that have been investigated suggest that a large set of broadly different structural strategies are available. The number of scaffolds that exist, however, is far fewer than the number of functions that are needed. Thus, the total number of protein folds present in all of biology has been estimated at ~500–1000 (May *et al.,* 1994), while the number of unique proteins in the human genome alone is likely to range from ~60,000 to 100,000 (Alberts *et al.,* 1994). Whatever variations in those numbers one prefers, it is clear that there are many more protein functions than folds available to deliver them. This leads to the inescapable conclusion that nature has reengineered the same scaffolds over and over to provide the range of functions required by biology.

With the advent of the genome projects, it now becomes possible to explore how nature has achieved functional diversity from a limited set of structural types by examining structure-function relationships in protein superfamilies. Our ability to do this at a sophisticated level is currently limited by the relatively small number of fold classes that have been structurally characterized, but recent initiatives to determine many more three-dimensional structures can be expected to substantially address this deficit over the next few years (Sali, 1998).

This chapter describes how nature has reengineered some protein scaffolds for a range of functions by examining a small number of example superfamilies. The goal is to illustrate how organizing very distantly related proteins into superfamilies provides a useful context for understanding evolutionary protein engineering that cannot be achieved by even the most elegant analysis of single proteins, one at a time. In this chapter we describe three enzyme superfamilies, the enolase, vicinal oxygen chelate fold (VOC), and haloacid dehalogenase (HAD) superfamilies, which we believe effectively exemplify some of the different types of structural engineering that nature has employed over the course of evolution.

II. Defining Protein Superfamilies

A. *Classical Definitions vs. Definitions that Include Functional Characteristics*

Before embarking on an analysis of specific superfamilies, it is important to define what we mean by the term. One classical definition simply describes protein families as homologous proteins of >50% sequence identity and superfamilies as evolutionarily related proteins of <50% sequence identity (Li and Graur, 1991). Note that this definition uses sequence information alone to determine relationship. At 50% identity, related sequences can usually be creditably aligned and functional information is not generally required to infer homology. But homologous proteins in superfamilies may also exhibit percent sequence identities well below 50% and have diverged to the point that they mediate very different overall functions. As will be shown, this is the most interesting window for examination of protein reengineering through natural evolution. It is also the window in which sequence information alone may be insufficient to confirm divergent relationship. To increase the credibility of these assignments, we have incorporated functional information.

There is another important reason to include functional information in the definition of superfamilies containing highly divergent proteins. This is because our analysis focuses on important structurally conserved elements that can be explicitly associated with fundamental characteristics of function. By defining protein function in this way, we make the explicit assumption that protein superfamilies evolved to deliver function in a specific and unique manner. These structural strategies for delivery of function can be recognized most easily when the problem is viewed in terms of linked structure-function relationships.

B. *Enzymes as Model Systems for Superfamily Analysis*

In principle, describing protein superfamilies in terms of the associated structure-function paradigm can be applied to any class of proteins. In our initial analyses we have focused on enzymes rather than on other classes of proteins because the relationships between protein structure and function can be more easily identified and tested in enzymes than in classes such as transporters or structural proteins. This is because individual steps of enzymatic reactions can be mapped explicitly to specific elements of the associated structures, giving us the direct correlation between protein structure and function required. Breaking down enzyme function into such individual mechanistic steps is especially important for identifying the common elements of function that are conserved among the structures of very divergent enzymes. Overall function, on the other hand, is frequently not shared by all divergent members of a superfamily.

C. *Issues in Superfamily Analysis*

A simple protocol for superfamily analysis is given in Figure 1. Note that it begins with the identification and verification of structural similarities, using both sequence and, as available, three-dimensional structural information. It is critical that the analysis begins with sequence and structural analysis rather than function. This is because nature has frequently evolved multiple structural strategies for delivering similar functions. Thus, unique mappings between structure and function can only be obtained by starting with a set of structurally similar proteins. The converse, an assumption of common evolutionary origin for similar functions, does not hold.

It is equally important that the evaluation of initial assignment of sequences to a superfamily be made from structural rather than functional criteria. This avoids, for example, trivial but important errors from the use of keyword searches to identify members of a superfamily. It also avoids early bias in determining residues/motifs that are conserved across all members of a superfamily. Multiple alignment and motif analysis provide sensitive methods for structure-based confirmation of sequence similarity. Scoring systems that can be applied to such methods provide verification of the statistical significance associated with the comparisons.

For these reasons, it is important to focus on the most divergent set of superfamily members that can be identified. Although a variety of new methods have recently been developed for identification of distantly related protein sequences [see, for example, Psi-Blast (Altschul *et al.,* 1997), methods based on Hidden Markov Models such as SAMT98 (Karplus *et al.,* 1998), the Intermediate Sequence Search algorithm of Park *et al.,* (Park *et al.,* 1997), or the simple congruence method, Shotgun (Pegg and Babbitt, 1999)], confirmation of these relationships can be technically difficult. In some cases, three-dimensional structural information or experimental structure-function analysis will be required to pro-

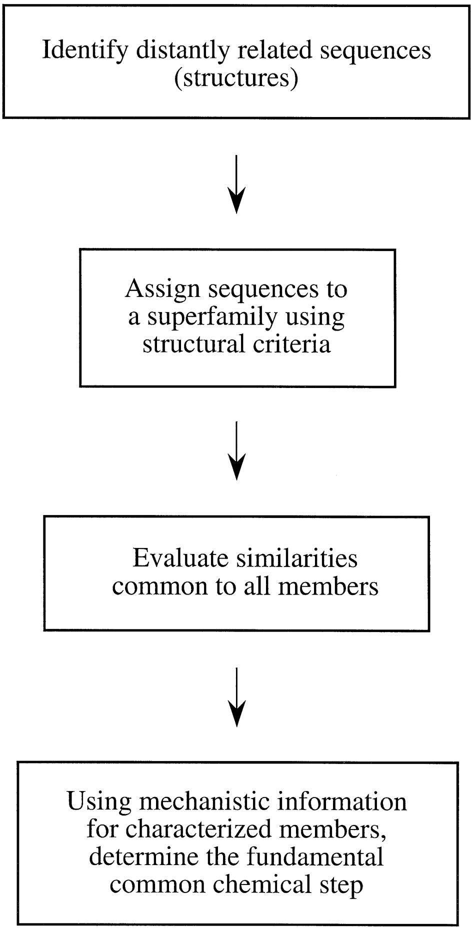

Identify distantly related sequences (structures)
Assign sequences to a superfamily using structural criteria
Evaluate similarities common to all members
Using mechanistic information for characterized members, determine the fundamental common chemical step

FIG. 1. Protocol for superfamily analysis.

vide convincing evidence of evolutionary relationship among the most distant of candidate homologs.

Once the set of superfamily members has been confirmed, the next step is to identify conserved sequence elements that can be associated with conserved functional steps among the member proteins. This will be most easily achieved for superfamilies in which one or more members have been mechanistically characterized and residues or motifs playing important chemical roles already identified. Here, inference based on analogy can be highly useful in predicting properties of uncharacterized proteins. This, generally, is one of the most common procedures in

Bioinformatics analysis and is routinely used to identify functions of closely related sequences from database searching. For more divergent proteins such as those discussed here, multiple alignment of homologs can be used to determine conserved motifs or residues. This information may be sufficient to infer important properties of poorly characterized proteins or of unknown reading frames. Even when little biological information about any of the homologs is available, patterns of conservation across divergent sets of superfamily members may be useful to form hypotheses about conserved function.

III. The Enolase Superfamily: A Paradigm for Understanding How Nature Evolves Enzymes for New Functions

We begin this discussion with an analysis of the enolase superfamily for several reasons. First, as one of the best-described superfamilies to date, it provides a good example of the usefulness of describing structure-function relationships in the superfamily context. Consistent with the criteria we have identified for effective superfamily analysis, it includes a wide range of highly divergent enzymes that catalyze sometimes very different overall reactions using substantially different substrates. Based on the similarities among six superfamily members that have so far been structurally characterized, all the superfamily members are predicted to have highly similar three-dimensional structures. Despite the functional differences among them, this conservation in structure appears to apply even to the identity and geometries of functionally important residues in the respective active sites. Finally, the enolase superfamily scaffold is an α/β barrel, one of the most common and best characterized fold types in all of biology (Farber and Petsko, 1990; Gerstein, 1997; Gerstein *et al.,* 2000). Thus, analysis of the manner in which the enolase superfamily evolved to deliver function may be useful for interpreting structure-function relationships in other superfamilies of α/β barrels.

A. *Early History: The Mandelate Racemase/Muconate Lactonizing Enzyme Story*

The enolase superfamily story started with the serendipitous discovery that two enzymes catalyzing very different overall reactions, mandelate racemase (MR) and muconate lactonizing enzyme (MLE), had virtually superimposable structures (Neidhart *et al.,* 1990). As shown in Figure 2, MR catalyzes the reversible racemization of mandelate, an aromatic substrate, while MLE catalyzes the equilibration of muconolactone with *cis,cis*-muconate. Given the substantial differences in these reactions, it

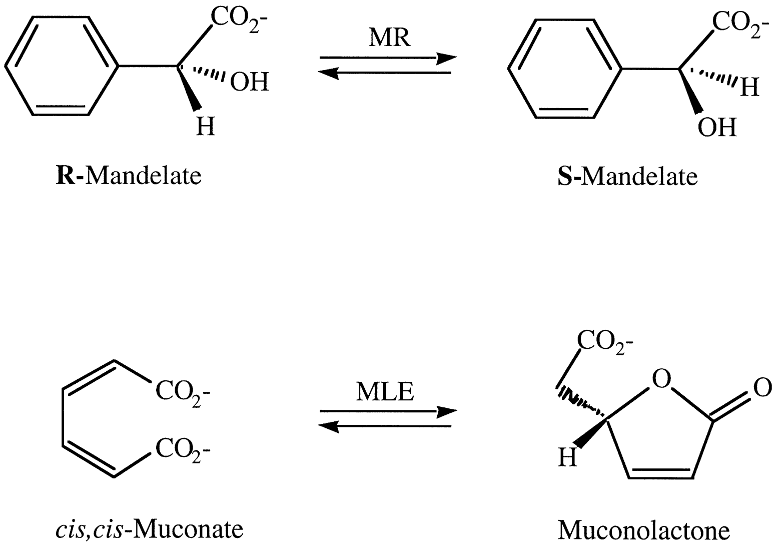


FIG. 2. Overall chemical reactions of MR and MLE.

was particularly surprising to find that even the active sites of the two enzymes have highly similar geometries (Fig. 3, see color insert), raising the question: How do two such dissimilar chemistries come to be mediated by such highly similar structures?

The critical insight for answering this question came from analysis, not of the overall reactions, but of the individual steps that make up the mechanisms of each. Here, two common elements of function could be identified. First, both MR and MLE are metal dependent enzymes, each using a divalent metal ion as an obligate cofactor in their mechanisms. Second, both enzymes initiate their reactions by abstraction of a proton α to a carboxylate group in their respective substrates leading to a common type of stabilized enolate intermediate. Not only were these fundamental functional characteristics common to both enzymes, the conserved functional groups in their active sites were precisely those that had been verified, through mechanistic studies of MR, to be responsible for these aspects of the chemical function (Landro *et al.,* 1991; Neidhart *et al.,* 1991; Powers *et al.,* 1991; Landro *et al.,* 1994; Kenyon *et al.,* 1995). Thus, the common fundamental mechanistic steps that both enzymes share are those that map explicitly to the conserved elements in their active sites and can be essentially "read off" the superposition shown in Figure 3.

Legends for Color Insert

FIG. 3. Superposition of the active sites of MR and MLE showing conserved active site residues.

FIG. 5. Alignment of MR and MLE with the first five homologs identified. The sequences of MR and MLE are shown in red. Numbering is that for the MR sequence. Conserved residues discussed in the text are shown in large size, black, bold lettering. Residues colored green designate a nonhomologous residue at that position that would not be expected to provide the requisite chemical capability associated with that conserved motif. (A) Sequence region containing the proton abstraction machinery for an S-substrate. (B) Sequence region containing the three metal binding ligands. The ''N'' residue at the 247 position in GlucD has been shown from crystallographic studies to perform as the third metal binding ligand in this protein even though it is not correctly aligned by multiple alignment algorithms (Gulick *et al.,* 1998). (C) Sequence region containing the proton abstraction machinery for an R-substrate (positions 270, 297) and the active site acidic residue involved in stabilization of the enolate intermediates predicted to be formed in all of the enzymes (position 317). Swiss Protein (SP) or Genbank (GB) database accession numbers for the sequences are: RspA, URF (*E. coli*): SP P38104; Spa2, URF (*Streptomyces ambofaciens*): SP P32426; GalD, galactonate dehydratase (*E. coli*): SP P31458 (this sequence includes an N-terminal region of an aldolase as explained in the text); MLE, muconate lactonizing enzyme I (*Pseudomonas putida*): GB, U12557 (the sequence used in the alignment differs slightly from U12557); MR, mandelate racemase (*Pseudomonas putida*): SP P11444; rTSa, URF (*Homo sapiens*): GB X67098; GlucD, glucarate dehydratase (*Pseudomonas putida*): SP P42206.

FIG. 8. Conserved motifs for some divergent proteins in the enolase superfamily. Residues known or predicted to be involved in proton abstraction are marked with a vertical arrow and residues acting as metal binding ligands are marked with an asterisk. Numbering is that for MR except for the K345 below the alignment in the purple-boxed motif, which is that of enolase. Yellow shading designates motifs described in the text. Green shading designates a nonhomologous residue at a conserved position that would not be expected to provide the requisite chemical capability associated with that motif. Because neither RspA nor Spa2 has been functionally or structurally characterized, ability to perform the relevant chemistry has not been ruled out and may be accessible using yet to be identified residues in the active site. The horizontal lines designate subgroupings related to structure-function relationships described in Babbitt *et al.* (Babbitt *et al.,* 1996). Proteins whose structures have been published (cited in the text) are designated with horizontal arrows next to the name abbreviations. The red-boxed motif contains the KXK motif associated with proton abstraction from an S-substrate. The blue-boxed motif contains the metal binding ligands. The purple-boxed motif contains residues known or predicted to be involved in proton abstraction from an R-substrate. Sequence information for MR, GalD, GlucD, RspA, Spa2, rTS (rTSA), MLE I (MLE) is given in the legend to Figure 5. Swiss Protein (SP) or Genbank (GB) database accession numbers for the remaining sequences are: MLE II, muconate lactonizing enzyme II (*Pseudomonas putida*): SP P11452; NAAAR, *N*-acyl amino acid racemase (*Amycolaptosis* sp.): GB D30738; βMAL, β-methylaspartate ammonia lyase (*Clostridium tetanomorphum*): SP Q05514; OSBSEc, *o*-succinylbenzoate synthase (*E. coli*): SP P29208;OSBSSsp, *o*-succinylbenzoate synthase (*Synechocystis sp.*): GB D64001; EnolSc, enolase (*Saccharomyces cerevisiae*): SP P00924; Enolhal, *enolase* (*Haloarcula marismortui*): SP P29201; CPEPS, carboxyphosphoenolpyruvate *synthase* (*Streptomyces hygroscopicus*): GB D37878; cMycBP, cMyc promoter binding protein (*Homo sapiens*): SP P22712.

FIG. 13. Superposition of α/β barrel domains of MR (white/red), MLE I (blue/red), enolase (yellow/red). Red coloring represents most similar regions among each pairwise comparison as described in the text.

This explicit mapping between structure and function for MR and MLE include the three metal binding ligands responsible for coordination to the divalent metal ions required for activity in MR (Mg^{2+}) and MLE (Mn^{2+}). These ligands, D195 and E221 in MR, are in virtually identical positions to those of their homologs, D198 and E224, in MLE. The functionally important carboxyl groups of the third metal binding ligands, E247 (MR) and D249 (MLE), are also oriented similarly. The proton abstraction machinery required to initiate each reaction can be identified on the either side of the active sites. On the left of the active sites shown in Figure 3, the conserved KXK motif can be associated with proton abstraction from the S-substrates, with K166 (MR) the S-specific base for S-mandelate and its homolog in MLE, K169, the S-specific base that initiates the reaction by abstraction of a proton from muconolactone. On the R-face of the active site, the pattern differs. Here, the proton abstraction machinery for MR is a His-Asp dyad, D270/H297, with H297 acting as the R-specific base. Since MLE performs the proton abstraction step only in the reverse direction, the Lys residue that superimposes with H297 in MR is expected to be only a spectator to the chemistry. As will be seen later, however, we propose that conserved active site lysines in this position in other superfamily members do act to abstract protons from R-substrates. As shown in Figure 3, both MR and MLE have another conserved residue in their active sites (E317 in MR, E327 in MLE) that is thought to be involved in stabilization of the enolate intermediate.

Based on these insights and the structure-function paradigm they define, the MR and MLE reactions can be rewritten to emphasize the common fundamental chemistry they share through structural conservation of their active sites (Fig. 4) (Petsko *et al.,* 1993).

B. Expansion of the Superfamily: Extending the Definition of the Structure-Function Paradigm

Following the discovery of the MR/MLE superfamily, the advent of the genome projects began to fill the public databases with enormous volumes of new sequence information. Not surprisingly, routine database searches identified new homologs of MR and MLE. The first five of these homologs discovered were all proteins of unknown function. From alignments of these proteins with MR and MLE, it was immediately clear that they shared many of the same conserved residues that had been identified as important to the functional steps common to both. Thus, multiple alignment of all seven proteins showed conservation of the three metal binding ligands associated with catalysis in MR and MLE as

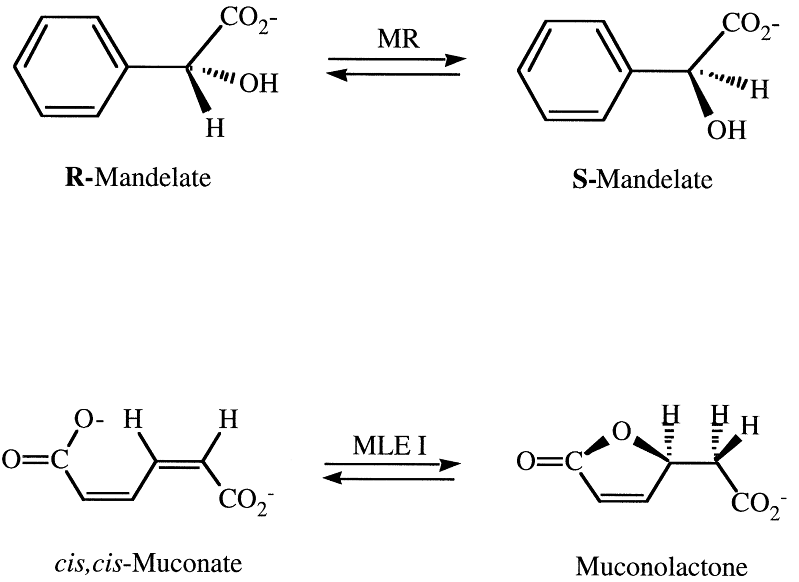


FIG. 4. Chemical reactions of MR and MLE written to emphasize the proton abstraction step.

well as the stabilizing Glu (E317 in MR) and one or both sets of proton abstraction machinery (Fig. 5, see color insert).

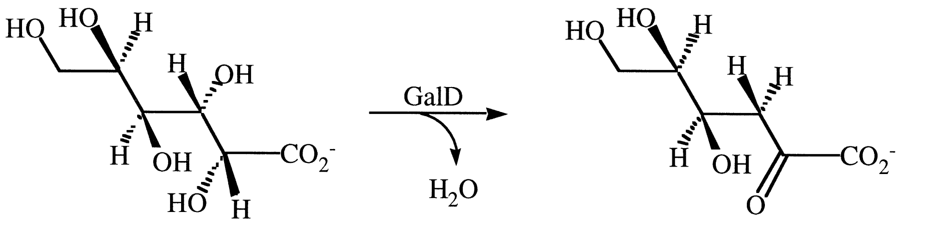
Among the most interesting of these new homologs was an unknown reading frame (URF) from the *E. coli* sequencing project, *orf*587, which appeared to contain the MR-like His/Asp dyad proton abstraction machinery for catalysis with an R-substrate, but was missing the K166 homolog responsible for proton abstraction from an S-substrate. If we consider MR the model for proton abstraction for either an R- or an S-substrate and MLE the model for proton abstraction of an S-substrate alone, then *orf*587 represents the third statement of the active site: proton abstraction from an R-substrate alone. The next goal, it appeared, was to determine whether the structure-function paradigm defined by MR and MLE could be used to infer functional properties of *orf*587.

1. *Using the Structure-Function Paradigm to Predict New Functions*

*a. Prediction of Function for An Unknown Reading Frame: orf*587. We reasoned that the paradigm that had been defined through the analysis of MR and MLE, if correct, could be used to infer two important properties of *orf*587: (1) that it catalyzes a reaction for a substrate whose stereochemical configuration is R- and (2) that it initiates the reaction by abstraction of a proton α to a carboxylate group in the substrate. From the partially completed *E. coli* sequencing project (Burland *et al.,* 1993), we were able to obtain the critical additional information necessary to identify the explicit function of this URF. That analysis showed the N-terminal third of *orf*587 to have some sequence similarity with sugar metabolizing enzymes, specifically bacterial aldolases. This was consistent with our analysis, which showed only the terminal two-thirds of the ORF to be homologous to MR and MLE. The sequencing project had also localized *orf*587 in the vicinity of minute 82 on the physical map of the *E. coli* chromosome. Reasoning that *orf*587 might represent one enzyme in a pathway associated with sugar metabolism, we searched near minute 82 of the genetic map of *E. coli* for such a pathway that would include an enzyme whose chemistry was consistent with our hypothesis.

We found such a pathway, encoded by the *dgo* operon, which is responsible for metabolism of the acid sugar galactonate (Babbitt *et al.,* 1995). One of the genes in this operon, *dgo*D, encodes the enzyme galactonate dehydratase (GalD), which exhibits the properties predicted by our analysis, initiation of the reaction by abstraction of an α-proton of an R-substrate, D-galactonate. Subsequent experimental confirmation of the predicted GalD activity also revealed that *orf*587 had been mistranslated due to a sequencing error that resulted in combination of a part of the 2-oxo-3-deoxygalactonate 6-phosphate aldolase gene, *dgo*A, and *dgo*D as one ORF, *orf*587.

The GalD reaction, drawn to emphasize the proton abstraction step common to members of the superfamily, is shown in Figure 6. From comparison of the GalD reaction with those of MR and MLE (Fig. 4), we suggest that it would have been virtually impossible to identify GalD without the conceptual framework provided by the earlier definition of the MR/MLE superfamily.



(*D*)-Galactonate

2-Keto-3-deoxy-(*D*)-galactonate

FIG. 6. Chemical reaction of GalD.

b. Prediction of Additional Functions for Glucarate Dehydratase. In addition to its usefulness for prediction of the function for *orf*587, the multiple alignment as shown in Figure 5 allowed prediction of additional functions for another superfamily member, glucarate dehydratase (GlucD). The GlucD function had been previously characterized (Jeffcoat *et al.,* 1969). Thus, GlucD was known to convert (D)-glucarate to 5-keto-4-deoxy-(D-glucarate. From this information, one would expect that the active site of GlucD would contain only a single general base for initiating catalysis via proton abstraction, that is, a homolog for K166 in MR. However, as shown in Figure 5, the alignment of GlucD with other known members of the superfamily depicts both a K166 homolog and a homolog for the H297/D270 dyad. Assuming that all members of the MR/MLE superfamily employ a common catalytic strategy, this suggested an additional reaction for this enzyme, dehydration of (L)-idarate. Experiment and three-dimensional structural characterization subsequently confirmed this prediction and helped to interpret the additional finding that the enzyme can also catalyze the epimerization of the two substrates (Palmer and Gerlt, 1996; Gulick *et al.,* 1998; Palmer *et al.,* 1998). All three reactions are shown in Figure 7. Again, as in the prediction of the GalD function, these additional functions of GlucD would not have been discovered without the inferences made obvious from examination of this sequence in its superfamily context.

2. *Characteristics of Conservation and Variation Across Divergent Members*

As more sequence information has become available, many additional members of the superfamily have been identified. These include the enzyme, enolase, as described by Babbitt *et al.* (Babbitt *et al.,* 1996). It is evident from an analysis of this larger set of homologs that nature has used this specific α/β barrel scaffold, evolved for metal dependent catalysis involving an initial proton abstraction step, to elaborate many different functions. Although many of these proteins are highly divergent from each other, exhibiting <20% sequence identity even in the conserved α/β barrel domain, all can be identified as members of the superfamily by the patterns of conserved residues shown in Figure 8 (see color insert).

In addition to the sequence similarities, structural similarities are evident among the six divergent members of the superfamily that have been structurally characterized. These include MR, MLE, and GlucD, as previously cited, as well as enolase (Lebioda and Stec, 1988; Lebioda and Stec, 1991; Wedekind *et al.,* 1995; Larsen *et al.,* 1996). These, along with preliminary structural characterization of two other superfamily members, GalD (J. Clifton, S. Wieczorek, J. Gerlt, and G. Petsko, unpub-

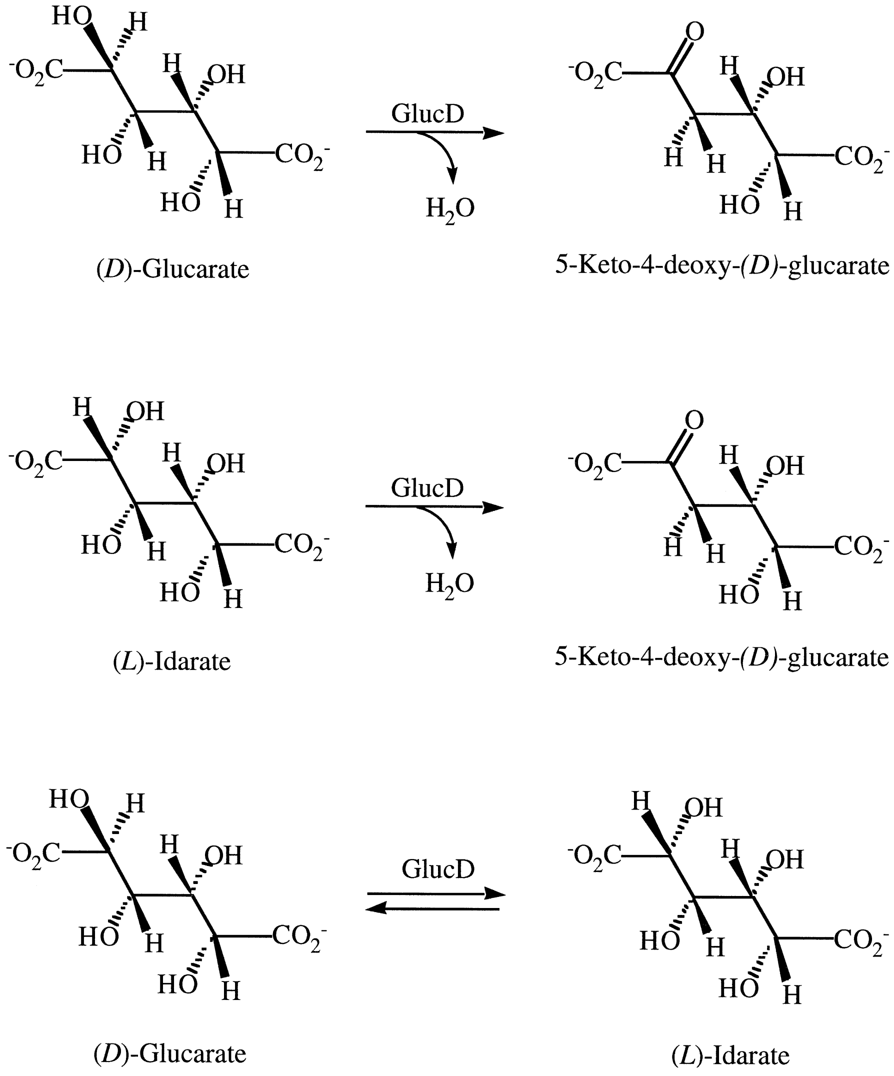


FIG. 7. Chemical reactions of GlucD.

lished) and *o*-succinylbenzoate synthase (OSBS) (T. Thompson, J. Garrett, J. Gerlt, and I. Rayment, unpublished), provide additional evidence that all the superfamily members are likely to share structural and functional similarities that are consistent with the structure-function paradigm we have described.

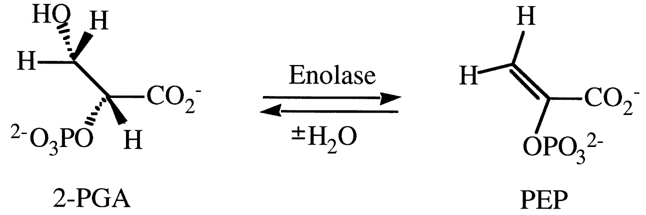


FIG. 9. Chemical reaction of enolase.

Addition of enolase to the superfamily is particularly important because it provides yet another major variation of the chemistries that can be mediated by the superfamily scaffold. As shown in Figure 9, the enolase reaction is also initiated by abstraction of a proton α to a carboxylate group in the substrate. Unlike any of the other chemistries that can be associated with either the MR or the MLE patterns, however, enolase exhibits substantially greater differences. Thus, the substrate of the enolase reaction is 2-phosphoglycerate, showing that phosphorylated substrates can be included in the range of substrates that the scaffold can accommodate. Enolase has two conserved metal ions, unlike other previously characterized members of the superfamily.

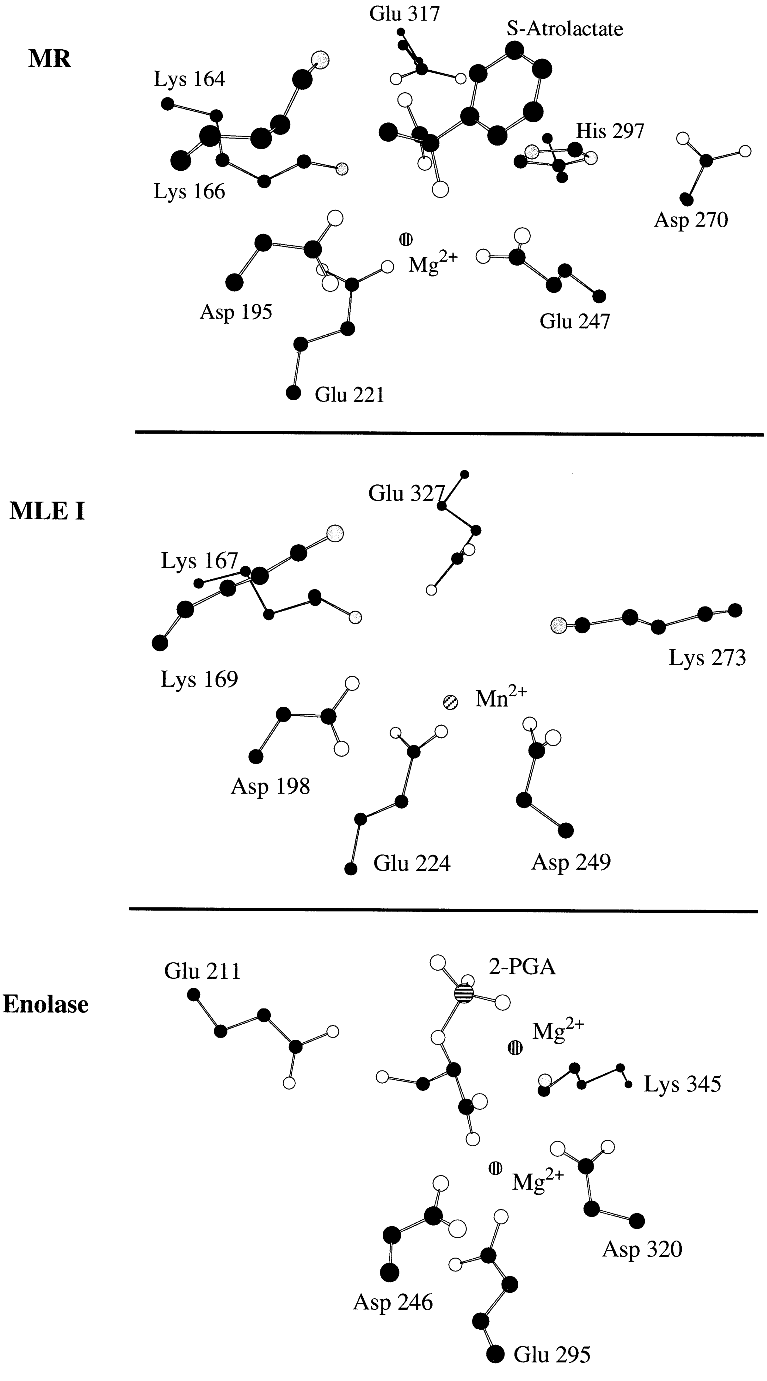
Despite the implications of these differences for substantial variations of the overall mechanisms, superposition of enolase with MR and MLE shows great similarity, even though the α/β barrel domain in enolase that contains the active site is only ~13% identical in sequence to either (Babbitt *et al.,* 1996). This similarity extends, as expected, to the geometries of the residues in the active site associated with the common chemistry of the superfamily, as shown in Figure 10. Enolase has only one set of proton abstraction machinery, however, a homolog for K273 in MLE, consistent with available mechanistic information (Poyner *et al.,* 1996; Reed *et al.,* 1997). This active site base is identified as K345 in Figure 8.

3. Another Variation on the Theme: Does N-acyl Amino Acid Racemase Represent Evolution in Action?

Of the individual members of the enolase superfamily so far determined to exhibit different overall functions, only one, *N*-acyl amino acid racemase (NAAAR) (Tokuyama and Hatano, 1995a; Tokuyama and

→

FIG. 10. Comparison of the active sites of MR (*P. putida*), MLE I (*P. putida*), and enolase (*S. cerevisae*). From Babbitt *et al.* (1996, Figure 1, p. 16490).


MR
Glu 317
S-Atrolactate
Lys 164
His 297
Lys 166
Asp 270
Mg^{2+}
Asp 195
Glu 247
Glu 221
MLE I
Glu 327
Lys 167
Lys 273
Lys 169
Mn^{2+}
Asp 198
Glu 224
Asp 249
Enolase
Glu 211
2-PGA
Mg^{2+}
Lys 345
Mg^{2+}
Asp 320
Asp 246
Glu 295

Hatano, 1995b) from *Amycolaptosis* sp. is known to catalyze more than one different chemical reaction using a substantially different substrate (Palmer *et al.,* 1999). Investigation of this catalytic flexibility in the context of the enolase superfamily raises the question of whether this enzyme may represent an example of nature's present-day reengineering of the superfamily scaffold for an entirely new function. Other examples of catalytically promiscuous enzymes from other superfamilies have been observed, as reviewed by O'Brien and Herschlag (O'Brien and Herschlag, 1999).

Like MR, NAAAR represents another racemase activity in the enolase superfamily. As would be expected, this enzyme contains machinery for initiation of its chemical reaction by abstraction of the α-proton from either an R- or an S-substrate. Thus, NAAAR possesses the S-specific base motif, KXK, as well an R-specific base, this time a Lys, in contrast to the His/Asp dyad in MR (Fig. 8). For this reason, NAAAR was originally assigned to the MLE subgroup, the structurally characterized member of the superfamily to which it is most similar (29% identical to the MLE I of *P. putida*) and which also has a Lys residue in the R-specific base position. Unlike any other members of the superfamily, however, NAAAR is a remarkably inefficient enzyme ($k_{cat}/K_m = 3.7 \times 10^2$ M^{-1} s^{-1}).

Subsequent to the assignment of NAAAR to the enolase superfamily, routine database searches revealed a new homolog in *B. subtilis* that is 43% identical to NAAAR. Analysis of this gene in its operon context revealed it likely to be an *o*-succinylbenzoate synthase (OSBS), an enzyme whose orthologs in several species had also been determined to belong to the enolase superfamily. The NAAAR and OSBS reactions are shown in Figure 11. The high degree of sequence identity between the putative OSBS from *B. subtilis* and NAAAR suggested that NAAAR, in turn, might also possess OSBS activity. Experimental investigation of this hypothesis showed that not only was NAAAR capable of catalyzing the OSBS reaction ($k_{cat}/K_m = 2.57 \times 10^5$ M^{-1} s^{-1}), it was much better at the OSBS reaction than at its originally characterized chemistry (Palmer *et al.,* 1999)! Further investigation showed that other OSBS enzymes from several species were unable to catalyze the NAAAR reaction. Taken together, the evidence suggested that the true physiological role of NAAAR is, in fact, the OSBS reaction, with the racemization chemistry a fortuitous side reaction of this particular enzyme from *Amycolaptosis.*

Although a convincing explanation for the presence of the low efficiency racemization chemistry remains a subject of speculation, these results suggest that NAAAR/OSBS represents a unique engineering solution within the enolase superfamily. (It will be interesting to see whether other superfamily members are also capable of equally surprising addi-

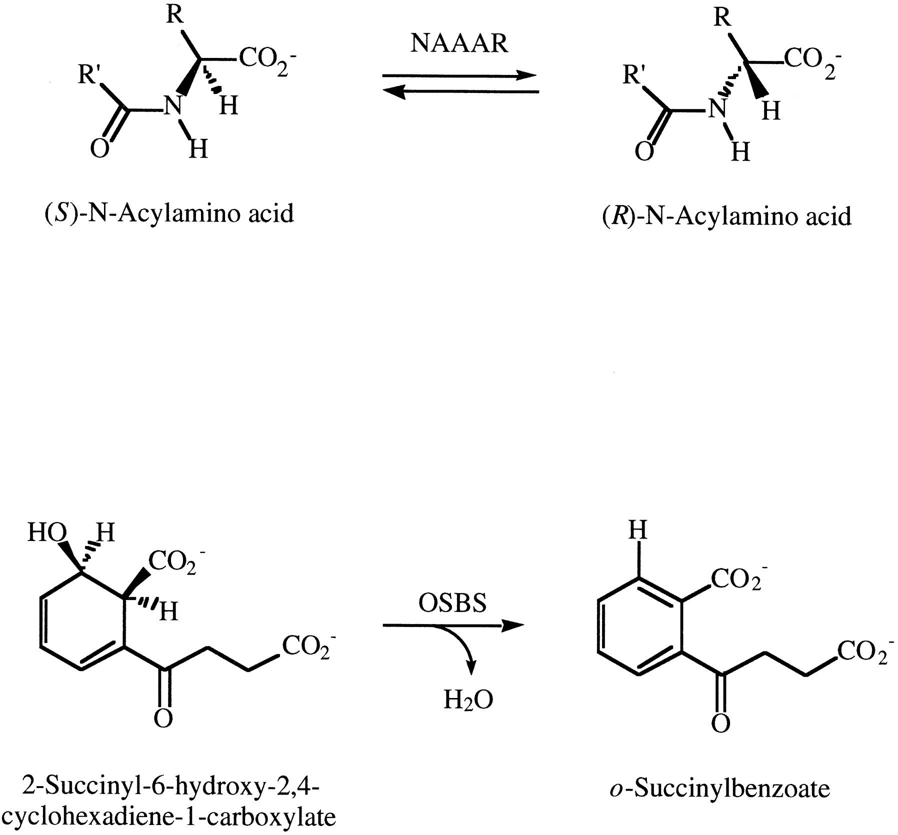


Fig. 11. Chemical reactions of NAAAR and OSBS.

tional chemistries, however.) Perhaps this is how nature engineers new functions into the scaffold. If so, two salient points are obvious. First, comparison of the NAAAR reaction with the OSBS reaction (Fig. 11) shows them to be substantially different with regard to substrate size and type as well as the overall chemical reactions they catalyze. Moreover, it appears to us that it would be extremely unlikely that human engineers would predict that an OSBS could perform NAAAR chemistry or *vice versa.* It seems, at least from this example, that we understand very little about how nature chooses template structures and chemistries for reengineering. Second, despite the remarkable differences between the OSBS and NAAAR reactions, both are entirely consistent with the structure-function paradigm we have described. Thus, while it appears that nature can accommodate even the very different OSBS and NAAAR reactions with one structure, it has performed that feat using a rather rigidly conserved structure-function paradigm. Whether these themes will continue to apply as new members of the superfamily are identified remains to be seen.

4. Nature's Engineering Principles for the Enolase Superfamily

a. Primary Constraints in Nature's Reengineering the Enolase Superfamily for Many Functions. Given the broad range of substrates and products represented in the enolase superfamily, it appears quite clear that nature conscripted this scaffold for reengineering based on the common fundamental chemical step that it catalyzes rather than on its ability to bind similar substrates. This notion contradicts a primary doctrine for enzyme evolution initially proposed by Horowitz, that binding a common substrate is the critical constraint in the evolution of new functions from a particular structural scaffold (Horowitz, 1945; Horowitz, 1965). Although it is likely that across the breadth of the protein universe both paradigms obtain, depending on the structural and functional characteristics of each superfamily, it is clear that there are a growing number of examples of superfamily evolution in which the conservation of chemical capabilities is a dominant constraint. (For short reviews of some of these superfamilies see Babbitt and Gerlt, 1997; Gerlt and Babbitt, 1998). Use of this notion as a critical principle in interpreting the structure/function relationships of the enolase superfamily has allowed us to correctly predict function and mechanism for a number of the member proteins. In addition, even for those members of the superfamily for which we are still unable to identify the precise substrates or overall enzymatic reactions, the powerful conceptual framework implicit in the structure-function paradigm we have described allows us to predict important active site residues from sequence alignments alone (see Babbitt *et al.* 1996).

Recently, Hwang *et al.* have performed conceptually similar predictions using three-dimensional structural information (Hwang *et al.,* 1999). Here, structural similarities between a protein of unknown function and others available in the public databases were used to infer the general function of the URF as a new nucleotide triphosphosphatase. It is expected that as many new structures of URFs are solved over the next few years, the conceptual approaches described here will be useful in interpreting characteristics of such proteins even in the absence of a clear understanding of overall function or identification of specific substrates and products.

b. General Structural Characteristics That Describe the Manner in Which the Enolase Superfamily Scaffold Delivers Function. The proteins of the enolase superfamily, like all α/β barrel enzymes (Farber and Petsko, 1990), deliver function in an active site effectively centered in the spherical locus of the barrel. Emanating from the beta sheets that form the barrel or the loops connecting them are functionally important residues poised to catalyze a chemical reaction involving an appropriately positioned substrate bound in the barrel. As shown in Figure 12 and described previously for the enolase superfamily (Babbitt and Gerlt, 1997; Hasson *et al.,* 1998), the evolutionary advantage offered by such an arrangement may help to explain the predilection nature apparently has for using the α/β barrel as an engineering template of choice: Evolution of new chemistries can be obtained essentially by point mutations of functionally important residues ringing the substrate(s) from any of the eight possible octant positions surrounding the barrel. These changes, accompanied by alterations in the active site to accommodate new substrates, intermediates, and products provide, to a simplified first approximation, the architectural schema used in the α/β barrels for the evolution of new enzymes. Moreover, these evolutionary ''hot spots,'' because they are presented on individual secondary structural units, can evolve more or less independently, without obvious consequences for distortion in

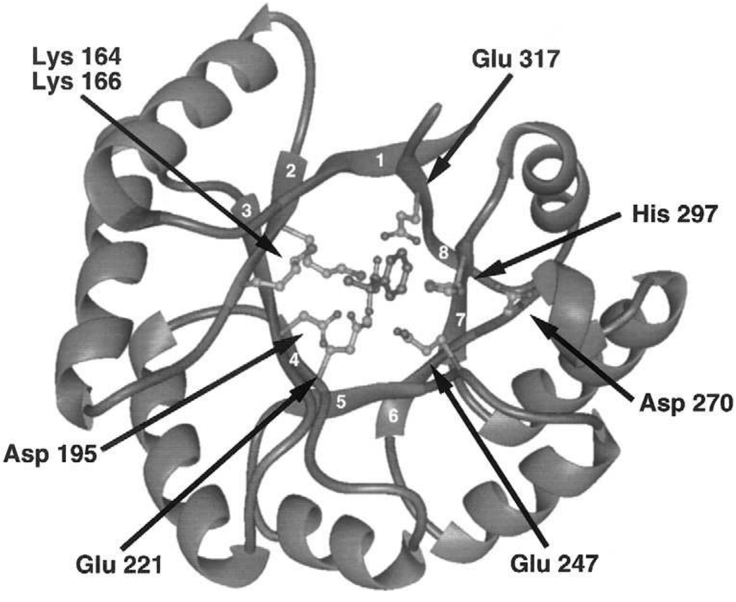


FIG. 12. α/β barrel domain of MR (based on Protein Data Bank entry 1 mnr). Important active site residues and the associated secondary structure elements are labeled and designated with arrows. From Babbitt and Gerlt (1997, Figure 1, p. 30592).

nearby beta sheets. The result, as interpreted for the enolase superfamily, allows for generation of new catalytic functions while retaining the common fundamental chemical step conserved in all members.

Support for this general interpretation is provided from a simple superposition experiment. All possible pairs of the MR, MLE, and enolase α/β barrels were superimposed and the RMSDs of all α-carbons calculated. Those RMSDs that fell below a defined threshold for all superpositions were then colored red to visualize where all three structures, relative to each other, showed the greatest similarity. One such visualization is shown in Figure 13 (see color insert). As can be seen in this figure, most of the variation (regions not colored red) among all three structures occurs in two outer loop regions and in the upper right quadrant of the structures. This relatively variable region is the same for all three proteins, indicating that nature altered this region most extensively in each protein relative to the others. This region is near the attachment sites for the N-terminal domains in each protein. These domains have been shown from crystallographic studies to be involved in interactions with the substrate. This is precisely the region that would be most expected to require retooling to accommodate the range of substrate sizes and chemical moieties exhibited within the superfamily. Conversely, the large red-colored region in Figure 13 represents the regions of these proteins that have been changed the least relative to each other. The center of this region, opposite the multicolored region, is the ''bottom'' of the α/β barrel, where the highly conserved metal binding ligands reside. In other words, the mapping of conservation shown in red is consistent with the notion that the fundamental chemistry associated with metal-assisted catalysis has been least disturbed during nature's reengineering of the ancestral scaffold.

Although it is tempting to suggest that the simple discrimination of conserved versus variable regions illustrated in Figure 13 (see color insert) could be used to help direct reengineering efforts for this superfamily in the laboratory, the complexities and surprises associated with our investigation of these proteins also suggest caution. Nevertheless, these observations are consistent with the accumulated sequence, structural, and functional evidence and suggest that the members of this superfamily have been retooled by relatively simple mechanisms inherent in the architecture of the α/β barrel structure. As ongoing DNA shuffling experiments in reengineering of members of this superfamily are completed, it will be interesting to see how closely the results match the model presented here.

5. How Do Other Superfamilies of the α/β Barrel Fold Fit into This Model?

Many other α/β barrel enzymes have been characterized both functionally and structurally and α/β barrels now represent some twenty-two superfamilies in the SCOP (Structural Classification of Proteins) classification of known three-dimensional protein structures (Murzin *et al.,* 1995). Among the α/β barrels, we and others have analyzed the N-acetylneuraminate lyase superfamily (Babbitt and Gerlt, 1997; Lawrence *et al.,* 1997) and Holm and Sander have described the amidohydrolase superfamily (Holm and Sander, 1997). In these superfamilies, the general paradigm described for the enolase superfamily obtains: a fundamental partial reaction in all the divergent members of the superfamilies can be explicitly mapped to conserved elements of the structures and used to interpret structure-function relationships across the superfamily.

Although these observations suggest that we have attained some important insights into how nature has reengineered this fold class for new functions, they do not answer the question whether all of the α/β barrels diverged from a single or at most a few common ancestors. If divergent, the important unanswered question becomes: How did entirely different chemistries, without even a partial reaction in common, evolve from an ancestral structure? This question has been debated extensively in the literature and will not be discussed in this review, yet it remains unclear how the structure-function paradigm we have described fits in to this larger picture. Perhaps the simplicity of the way in which function is delivered in α/β barrels allows, through point mutations and additional modest changes, entirely new chemistries to evolve. Perhaps the Horowitz model, in which new functions are conscripted on the basis of substrate-binding specificity, obtains in these furthest reaches of evolutionary divergence. Whatever the answer(s), it seems likely that we will have to await many new structures and superfamily analyses to address more fully the relationships among superfamilies of the α/β barrel fold. We expect that new exercises in combinatorial protein engineering will offer some additional explanations as well.

IV. Studies of Other Superfamilies and Fold Classes

Given the conclusions generated from the enolase superfamily analysis, it would be valuable to determine whether the principles we suggest are relevant to other fold classes besides the α/β barrels. To date, even including α/β barrel proteins, only a few studies have been completed that focus on the explicit structure-function correlations among highly divergent members of enzyme superfamilies. Some examples of such studies are given in Table I. In each case, a fundamental chemical step, rather than binding of similar substrates, is the functional characteristic that correlates with conserved structural features. For each of these superfamilies these fundamental functional characteristics, along with the relevant references, are also included in the table.

TABLE I
Some Examples of Enzyme Superfamilies Whose Members Mediate a Common Fundamental Chemical Step

Superfamily name (reference)	Fundamental common chemistry
Enolase (Babbitt *et al.,* 1996)	Stabilization of enolate anions generated from abstraction of a proton α to a carboxylate
Haloacid dehalogenase (Baker *et al.,* 1998)	Hydrolysis, phosphoryl group transfer via hydrolytic nucleophilic substitution
Vicinal oxygen chelate fold (Bergdoll *et al.,* 1998)	Stabilization of diverse oxyanion intermediates via metal-assisted catalysis
N-acetylneuraminate lyase (Lawrence *et al.,* 1997)	Schiff base dependent formation of an ''electron sink''
Amidohydrolase (Holm and Sander, 1997)	Hydrolysis of amides, ureas, phosphotriesters, triazines
Crotonase (Gerlt and Babbitt, 1998)	Stabilization of oxyanion intermediates derived from thioesters

Although a full exposition of these additional superfamilies is beyond the scope of this discussion, a short description of two, the vicinal oxygen chelate fold (VOC) superfamily and the haloacid dehalogenase (HAD) superfamily, are included. Neither of these superfamilies belongs to the α/β barrel fold class and a discussion of the way in which each delivers function provides an important contrast to our discussion of the enolase superfamily.

A. *The VOC Superfamily: Reengineering the Scaffold through Wholesale Domain Shuffling*

The VOC superfamily, like the enolase superfamily, includes proteins whose substrates, products, and overall chemical reactions are very different. These include the large class of extradiol dioxygenases [see Eltis and Bolin for a good review (Eltis and Bolin, 1996)], glutathione-dependent lactoylglutathione lyases (glyoxalase I) and antibiotic-resistant proteins such as fosfomycin resistance protein (Bernat *et al.,* 1997; Laughlin *et al.,* 1998; Bernat *et al.,* 1999), and methylmalonyl-CoA epimerase (T. Haller and J. Gerlt, unpublished). Although each type of reaction is dependent on a different metal ion, the structure/function paradigm can be generally described in terms of metal-assisted catalysis in which the metal provides electrophilic assistance through chelation of substrates/intermediates leading to stabilization of an oxyanion intermediate (Babbitt and Gerlt, 1997). This is reminiscent of our description of the enolase superfamily insofar as a common fundamental functional step can be identified in all the divergent members that maps explicitly to the similarities in their active sites.

The architectures of these proteins are very different from those of the α/β barrels, however, and this has dictated a very different set of structural possibilities for delivery of function than is seen in the enolase superfamily. Where we can describe the primary changes in reengineering the enolase superfamily in terms of single point mutations to obtain new chemistries and active site variation to accommodate new substrates, the overall strategy that nature has used in reengineering the VOC scaffold is more complex. As has been elegantly described by Bergdoll *et al.* (Bergdoll *et al.,* 1998), analysis of the three-dimensional structures available for different members of the VOC superfamily shows that nature has generated new structures and functions by shuffling whole domains through multiple gene duplication and fusion events. In all cases, the critical metal-binding residues are conserved and can be used as primary criteria for assigning new sequences to the superfamily. Interestingly, recent analysis of a new subgroup within the extradiol dioxygenase family suggests that this paradigm has also been used for catalysis with a substrate in which the OH- groups are *para,* obviating the possibility for chelation of the metal ion (Xu *et al.,* 1999). Even in this case, however, a reasonable mechanism can be written that is consistent with the superfamily paradigm.

B. *The HAD Superfamily: A Common Chemistry for Hire?*

The HAD superfamily, again, like the enolase and VOC superfamilies, represents highly divergent enzymes catalyzing many different functions. Again, the common active site machinery can be correlated with a common functional step in all the member mechanisms. Of the superfamilies represented in Table I, the sequences of the HAD superfamily are the most divergent. These include the haloacid dehalogenase function, many phosphatases, phosphonatases (in which the cleavage reaction involves a P–C bond rather than the P–O bond normally associated with phosphate esters), β-phosphoglucomutase, and a family of P-type ATPases (Koonin and Tatusov, 1994; Aravind *et al.,* 1998; Baker *et al.,* 1998).

Comparison of available three-dimensional structures for two haloacid dehalogenases (Hisano *et al.,* 1996; Ridder *et al.,* 1997) with that of a recently solved structure for a phosphonatase (M. Morais and K. N. Allen, unpublished) shows them to be superimposable, with substantial active site similarities associated with an aspartic acid that is rigorously conserved in all the member sequences. As reviewed by Baker *et al.* (Baker *et al.,* 1998), mechanistic studies of several of the superfamily members show that this aspartic acid functions in nucleophilic catalysis. Thus, it is the central player in the structure-function paradigm that can be described for this superfamily—that is, a general structural strategy for hydrolytic nucleophilic substitution.

The architectural scheme by which the members of the HAD superfamily deliver function differs substantially from that described above for either the enolase or VOC superfamilies. Representing a new and unique fold type, the HAD superfamily members can apparently serve as integral modules in varied and complex larger protein structures. As shown in Figure 14, superfamily members are observed as single soluble proteins as represented by HAD, phosphonatase, and other proteins of unknown function, as membrane-bound proteins, as in phosphoserine (PSP) phosphatase, as a domain in bifunctional proteins such as the histidinol phosphate phosphatase or as a domain of unknown function in the soluble epoxide hydrolases from animals, or as the ATP-hydrolyzing catalytic domain of some P-type ATPases. The ability of this scaffold to deliver catalysis via a conserved chemical step in such a broad range of structural contexts provides a remarkable example of the engineering versatility accessible to nature. As more members of this superfamily are identified and characterized, it will be interesting to see how this built-in versatility for adaptive localization can be exploited by protein engineers in the laboratory.

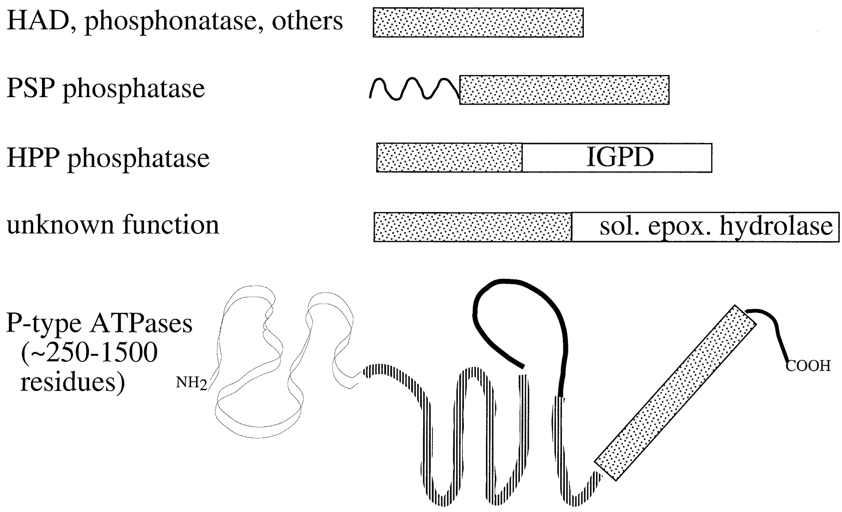


FIG. 14. Localization of homologous modules in the HAD superfamily in representative larger protein contexts as described in the text. Gray-stippled regions represent the HAD superfamily homologs.

C. Implications of Superfamily Analysis for Current Definitions of Enzyme Function

Analysis of protein superfamilies through explicit mapping of common functional characteristics to conserved structural elements among very distantly related homologs provides a valuable conceptual approach to connecting sequence, structure, and function in ways that are important for understanding both natural and laboratory-based protein engineering. From this perspective, the approach offers some progress in achieving a more systematic and structurally contextual description of enzyme function than has previously been available. This is because a definition of enzyme function in the superfamily context more precisely reflects the physical reality of protein evolution—that protein function evolves through changes in protein structure—than do earlier definitions that were developed without benefit of structural information. An example of the confusion that arises in describing enzymes only in terms of their functional characteristics is evidenced in the way the E.C. system (Bairoch, 1999) designates the members of the enolase, VOC, and HAD superfamilies. As shown in this review, the enzymes of these superfamilies can all be described in terms of common, correlated structural and functional characteristics. Yet these enzymes are assigned very different E.C. numbers, as shown in Table II. This conundrum arises from the history of the development of the E.C. system, whose conceptual ground-

TABLE II
E.C. Classifications for Some Enzymes in the Enolase, VOC, and HAD Superfamilies

Enzyme name	E.C. number
Enolase superfamily	
Galactonate dehydratase	4.2.1.6
Enolase	4.2.1.11
Glucarate dehydratase	4.2.1.40
Mandelate Racemase	5.1.2.3
Muconate Lactonizing Enzyme	5.5.1.1
VOC Superfamily	
Extradiol Dioxygenases	1.13.11.x
Lactoylglutathione Lyase	4.4.1.5
HAD Superfamily	
Phosphoserine phosphatase	3.1.3.3
Histidinol phosphatase	3.1.3.15
Phosphonatase	3.11.1.1
2-Haloacid dehalogenase	3.8.1.2
β-Phosphoglucomutase	5.4.2.6

work was established prior to the availability of structural information. As a result, the E.C. system's focus on the overall chemical reactions of enzymes does not map well to structurally defined clusters of related proteins. Although we expect that the E.C. system will continue to be useful for categorization of the substrates and overall chemical reactions of enzymes, it is not an appropriate database for generating correlations between protein structure and function, at least when viewed from the superfamily level.

V. Conclusions

Although all the superfamilies that we have examined to date show that chemistry is the dominant constraint in nature's reengineering of the respective scaffolds for new and different functions, examples also exist that may provide support for Horowitz' original proposals. For example, two proteins that function in histidine biosynthesis, phosphoribosylformimino-5-aminoimidazole carboxamide isomerase and imidazole glycerolphosphate synthase, catalyze successive steps in that pathway and are homologous (Fani *et al.,* 1998). As more enzyme superfamilies are characterized, it will be interesting to see how the evolutionary constraints associated with them partition between the two schemes proposed by us and by Horowitz. We suggest that knowing the answer to this issue for any particular superfamily will be valuable in providing appropriate direction for using that scaffold for protein engineering in the lab. We expect that this will be particularly true when new reactions or metabolic paths are sought to provide chemistries that are substantially different from any now known in nature.

Acknowledgments

The authors wish to thank Dr. Malin M. Young of Sandia National Labs for her contribution to Figure 13 and Dr. Karen Allen of Boston University for providing preliminary structural coordinates for the phosphonatase from *Salmonella typhimurium.* Molecular graphics images were produced using the MidasPlus program from the Computer Graphics Laboratory, University of California, San Francisco [supported by NIH 2-P41-RR01081, T.E. Ferrin, C.C. Huang, L.E. Jarvis and R. Langridge (1988). The MIDAS display system, *J. Mol. Graphics,* **6:**13–27.]

References

Alberts, B., D. Bray, J. Lewis, M. Raff, K. Roberts, and J. D. Watson (1994). *Molecular Biology of the Cell.* Garland Publishing, Inc., New York.

Altschul, S. F., T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman (1997). *Nuc. Acids. Res.* **25,** 3389–3402.

Aravind, L., M. Y. Galperin, and E. V. Koonin (1998). *TIBS* **23,** 127–129.

Babbitt, P. C., and J. A. Gerlt (1997). *J. Biol. Chem.* **272,** 30591–30594.

Babbitt, P. C., M. Hasson, J. E. Wedekind, D. J. Palmer, M. A. Lies, G. H. Reed, I. Rayment, D. Ringe, G. L. Kenyon, and J. A. Gerlt (1996). *Biochem.* **35,** 16489–16501.

Babbitt, P. C., G. T. Mrachko, M. S. Hasson, G. W. Huisman, R. Kolter, D. Ringe, G. A. Petsko, G. L. Kenyon, and J. A. Gerlt (1995). *Science* **267,** 1159–1161.

Bairoch, A. (1999). *Nuc. Acids. Res.* **27,** 310–311.

Baker, A. S., M. J. Ciocci, W. W. Metcalf, J. Kim, P. C. Babbitt, B. L. Wanner, B. M. Martin, and D. D. Dunaway-Mariano (1998). *Biochem.* **37,** 9305–9315.

Bergdoll, M., L. D. Eltis, A. D. Cameron, P. Dumas, and J. T. Bolin (1998). *Prot. Sci.* 7, 1661–1670.

Bernat, B. A., L. T. Laughlin, and R. N. Armstrong (1997). *Biochem.* **36,** 3050–3055.

Bernat, B. A., L. T. Laughlin, and R. N. Armstrong (1999). *Biochem.* **38,** 7462–7469.

Burland, V., G. Plunkett, D. L. Daniels, and F. R. Blattner (1993). *Genomics* **16,** 551–561.

Eltis, L. D., and J. T. Bolin (1996). *J. Bacter.* **178,** 5930–5937.

Fani, R., E. Mori, E. Tamburini, and A. Lazcano (1998). *Orig. Life & Evol. Biosph.* **28,** 555–570.

Farber, G. K., and G. A. Petsko (1990). *Trends Biochem. Sci.* **15,** 228–234.

Gerlt, J. A., and P. C. Babbitt (1998). *Curr. Op. Chem. Biol.* **2,** 607–612.

Gerstein, M. (1997). *J. Mol. Biol.* **274,** 562–576.

Gerstein, M., J. Lin, and H. Hegyi, Eds. (2000). *Protein folds in the worm genome.* Pacific Symposium on Biocomputing 2000. Honolulu, HI, World Scientific Press.

Gulick, A. M., D. J. Palmer, P. C. Babbitt, J. A. Gerlt, and I. Rayment (1998). *Biochem.* **37,** 14358–14368.

Hasson, M. S., I. S. Schlichting, J. Moulai, K. Taylor, W. Barrett, G. L. Kenyon, P. C. Babbitt, J. A. Gerlt, G. A. Petsko, and D. Ringe (1998). *Proc. Natl. Acad. Sci. USA* **95,** 10396–10401.

Hisano, T., Y. Hata, T. Fujii, J.-Q. Liu, T. Kurihara, N. Esaki, and K. Soda (1996). *J. Biol. Chem.* **271,** 20322–20330.

Holm, L., and C. Sander (1997). *Proteins: Struc. Funct. & Gen.* **28,** 72–82.

Horowitz, N. H. (1945). *Proc. Natl. Acad. Sci. USA* 31, 153–157.

Horowitz, N. H. (1965). In *Evolving Genes and Proteins* (V. Bryson and H. J. Vogel, eds), pp. 15–23. Academic Press, New York.

Hwang, K. Y., J. H. Chung, S. H. Kim, Y. S. Han, and Y. Cho (1999). *Nat. Struct. Biol.* **6,** 691–696.

Jeffcoat, R., H. Hassall, and S. Dagley (1969). *Biochem. J.* **115,** 969–997.

Karplus, K., C. Barrett, and R. Hughey (1998). *Bioinf.* **14,** 846–856.

Kenyon, G. L., J. A. Gerlt, G. A. Petsko, and J. W. Kozarich (1995). *Accounts of Chemical Research* **28,** 178–186.

Koonin, E. V., and R. L. Tatusov (1994). *J. Mol. Biol.* **244,** 125–132.

Landro, J. A., J. A. Gerlt, J. W. Kozarich, C. W. Koo, V. J. Shah, G. L. Kenyon, D. J. Neidhart, S. Fujita, and G. A. Petsko (1994). *Biochem.* **33,** 635–643.

Landro, J. A., A. T. Kallarakal, S. C. Ransom, J. A. Gerlt, J. W. Kozarich, D. J. Neidhart, and G. L. Kenyon (1991). *Biochem.* **30,** 9274–9281.

Larsen, T.M., J.E. Wedekind, I. Rayment, and G.H. Reed (1996). *Biochem.* **35,** 4349–4358.

Laughlin, L. T., B. A. Bernat, and R. N. Armstrong (1998). *Chem.-Biol. Interac.* 111–112, 41–50.

Lawrence, M. C., J. A. Barbosa, B. J. Smith, N. E. Hall, P. A. Pilling, H. C. Ooi, and S. M. Marcuccio (1997). *J. Mol. Biol.* **266,** 381–399.

Lebioda, L., and B. Stec (1988). *Nature* **333,** 683–686.

Lebioda, L., and B. Stec (1991). *Biochem.* **30,** 2817–2822.

Li, W.-H., and D. Graur (1991). *Fundamentals of Molecular Evolution.* Sinauer Assoc., Inc., Sunderland, MA.

May, A. C., M. S. Johnson, S. D. Rufino, H. Wako, Z.-Y. Zhu, R. Sowdhamini, N. Srinivasan, M. A. Rodionov, and T. L. Blundell (1994). *Phil. Trans. R. Soc. Lond. B* **334,** 373–381.

Murzin, A. G., S. E. Brenner, T. Hubbard, and C. Chothia (1995). *J. Mol. Biol.* **247,** 536–540.

Neidhart, D. J., P. L. Howell, G. A. Petsko, V. M. Powers, R. S. Li, G. L. Kenyon, and J. A. Gerlt (1991). *Biochem.* **30,** 9264–9273.

Neidhart, D. J., G. L. Kenyon, J. A. Gerlt, and G. A. Petsko (1990). *Nature* **347,** 692–693.

O'Brien, P. J., and D. Herschlag (1999). *Chem. and Biol.* **6,** R91–R105.

Palmer, D. R. J., and J. A. Gerlt (1996). *J. Am. Chem. Soc.* **118,** 10323–10324.

Palmer, D. R. J., J. B. Garrett, V. Sharma, R. Meganathan, P. C. Babbitt, and J. A. Gerlt (1999). *Biochem.* **38,** 4252–4258.

Pegg, S. C.-H. and P. C. Babbitt (1999). *Bioinformatics* **15,** 729–740.

Petsko, G. A., G. L. Kenyon, J. A. Gerlt, D. Ringe, and J. W. Kozarich (1993). *TIBS* **18,** 372–376.

Powers, V. M., C. W. Koo, G. L. Kenyon, J. A. Gerlt, and J. W. Kozarich (1991). *Biochem.* **30,** 9255–9263.

Poyner, R. R., L. T. Laughlin, G. A. Sowa, and G. H. Reed (1996). *Biochem.* **35,** 1692–1699.

Reed, G. H., R. R. Poyner, T. M. Larsen, J. E. Wedekind, and I. Rayment (1997). *Curr. Opin. Struct. Biol.* **6,** 736–743.

Ridder, I. S., H. J. Rozeboom, K. H. Kalk, D. B. Janssen, and B. W. Dijkstra (1997). *J. Biol. Chem.* **272,** 33015–33022.

Sali, A. (1998). *Nature Struct. Biol.* **5,** 1029–1032.

Tokuyama, S., and K. Hatano (1995a). *Appl. Microbiol. Biotechnol.* **42,** 884–889.

Tokuyama, S., and K. Hatano (1995b). *Appl. Microbiol. Biotechnol.* **42,** 853–859.

Wedekind, J. E., G. H. Reed, and I. Rayment (1995). *Biochem.* **34,** 4325–4330.

Xu, L., K. Resing, S. L. Lawson, P. C. Babbitt, and S. D. Copley (1999). *Biochem.* **38,** 7659–7669.

EVOLUTION OF PROTEIN FUNCTION BY DOMAIN SWAPPING

By MARC OSTERMEIER* and STEPHEN J. BENKOVIC†

*Department of Chemical Engineering, Johns Hopkins University, Baltimore, Maryland 21218, and †The Pennsylvania State University, Department of Chemistry, University Park, Pennsylvania 16802

I. Introduction

Obtaining proteins and enzymes with desired properties and activities is an important goal of biotechnology. However, even though recombinant DNA technology and molecular biology techniques allow the creation of a protein with any desired amino acid sequence, our current understanding of proteins falls short of that necessary for *de novo* design of protein function. Thus, two alternatives remain: (1) discovery of novel enzymes through scrutiny of natural resources or through recombinant genetic libraries (Short, 1997) and (2) remodeling of proteins already existing in nature. A balanced approach utilizing both strategies is likely to be the most successful, partly because new proteins discovered using (1) will create a larger database for performing (2).

0065-3233/00 $35.00

Protein remodeling by rational design utilizes our current understanding of how proteins function and, guided by computer modeling, identifies specific changes designed to impart the desired function onto the existing protein. Since our current understanding of enzymes is less than complete, combinatorial methods for protein remodeling that mimic evolutionary processes have become increasingly popular. Such remodeling of existing proteins can be performed by creating random libraries by methods such as error-prone PCR, cassette mutagenesis, DNA shuffling (Stemmer, 1994), molecular breeding (Crameri *et al.,* 1998), or by swapping elements of two or more existing enzymes to create hybrid enzymes. For the latter, the elements being exchanged may be individual residues, secondary structural elements, subdomains, domains, or whole proteins (i.e., fusion proteins) (Nixon *et al.,* 1998).

This chapter explores the use of large exchanges of structure (subdomains, domains, and whole proteins) to evolve new function in proteins. Both rationally designed and combinatorial exchanges are explored, as both mirror methods nature employs to evolve new function. Judicious application of both strategies will likely prove to be important for advances in the field.

II. Terminology

The terms ''domain'' and ''subdomain'' are used to describe proteins and fragments thereof. The term domain has been defined as ''a subregion of the polypeptide chain that is autonomous in the sense that it possesses all the characteristics of a complete globular protein'' (Schultz and Schirmer, 1979). The term subdomain usually refers to units smaller than a domain that can be described as having a particular function or structure (i.e., some rationale for being grouped together). What distinguishes the two from each other can often be confusing and the terms are frequently used interchangeably.

The problems in the use of the terms domain and subdomain result from their meaning being context dependent. What constitutes a subdomain for folding may not constitute a subdomain for a catalytic unit. For instance, the portion of a transcription factor responsible for binding DNA is most commonly called a ''DNA binding domain'' as it can bind DNA on it own, even though it lacks the complete function of a transcription factor because it is missing the activator domain. However, for an enzyme that can be divided into separate segments that are responsible for binding substrate and catalysis, these segments are commonly referred to as subdomains, even though the substrate binding subdomain may bind substrate on its own. The difference seems to come down to more intimate contact and functional coupling between subdomains than between domains. To avoid confusion, this chapter will use the term domain very loosely to refer to both domains and subdomains.

The term ''domain swapping'' refers to genetic rearrangement of proteins by combining pieces of genes that code for domains or subdomains. This is distinguished from the term ''3D domain swapping'' that describes the process in which one domain of a multidomain protein breaks its noncovalent bonds with the other domains and is replaced by the same domain of an identical protein chain (Schlunegger *et al.,* 1997).

III. Evolution of Proteins in Nature by Domain Swapping

The study and understanding of the natural evolution of protein function clearly has important implications for the design of *in vitro* evolution strategies. Because nature is blind to the rules and relationships of sequence, structure, and function, strategies that seem most applicable are combinatorial methods. Sequence and structural space are enormously large yet likely to be sparse in function. Nature must explore these spaces in a manner that preferentially examines areas relatively high in function. Thus, an understanding of how proteins evolve is an important step in rationally designing combinatorial protein engineering strategies.

It is generally accepted that nature evolves proteins for novel function by redesigning existing protein frameworks. This is supported by a number of lines of evidence: (1) the occurrence of gene duplications and recombination of genes and gene fragments is well established, (2) the relatedness of interspecies homologs observed by sequence and structural comparisons, (3) the distribution of motifs related by sequence and structure among disparate proteins, (4) the limited number of protein folds, with similar folds found in functionally dissimilar proteins, and (5) the deduction that sequence space is largely empty of function; hence, the *de novo* design of new proteins is too restrictive and inefficient a strategy to have evolved the current diversity and complexities of proteins, given the age of life on earth.

The creation of new function depends on minor and major changes in the sequence of existing genes. Minor changes include point mutations, deletions, and insertions of a few amino acids. Such minor changes can produce no effect, totally inactivate a protein, or lead to small or large changes in functionality. However, major changes in the sequences of genes through insertions of long sequences, tandem duplication, and circular permutations provide opportunities for major structural rearrangements and the evolution of entirely novel function. Major changes allow a protein to escape a local energy minimum and activate, in combination with other sequences, latent functionality. The importance of domain swapping for evolution is clearly illustrated by the evolution of *Escherichia coli* from *Salmonella* (Lawrence, 1997). None of the phenotypic differences between the two species can be attributed to point mutations. All the phenotypic changes have arisen by domain swapping.

The following sections explore nature's use of domain swapping to evolve new function. These include the formation of multifunctional proteins, tandem duplication, domain recruitment, and cicular permutation (Fig. 1). The evolution of several enzymes in the purine (Fig. 2) and pyrimidine (Fig. 3) *de novo* biosynthetic pathways, as well as other enzymes, are discussed as illustrative examples.

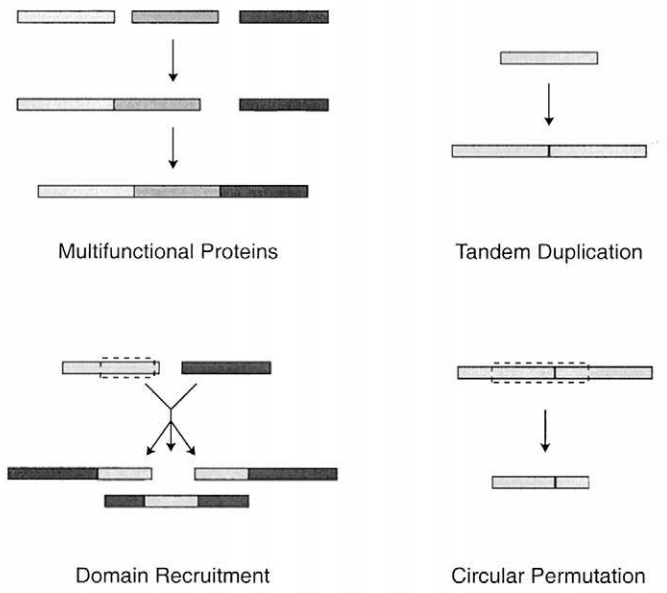


FIG. 1. Schematics of evolutionary mechanisms of domain swapping in nature. Multifunctional proteins arise from the fusion of the genes coding for individual enzymes. Often the individual domains of multifunctional proteins catalyze successive steps in metabolic pathways. In tandem duplication, a gene is duplicated and the 3′ end of one copy is fused in-frame to the 5′ end of the second copy. In domain recruitment, a functional unit (whole gene or gene fragment) from one gene is either inserted within or fused to an end of a second gene. Circular permuted genes are believed to arise via tandem duplication followed by introduction of new start and stop codons (Ponting *et al.,* 1995).

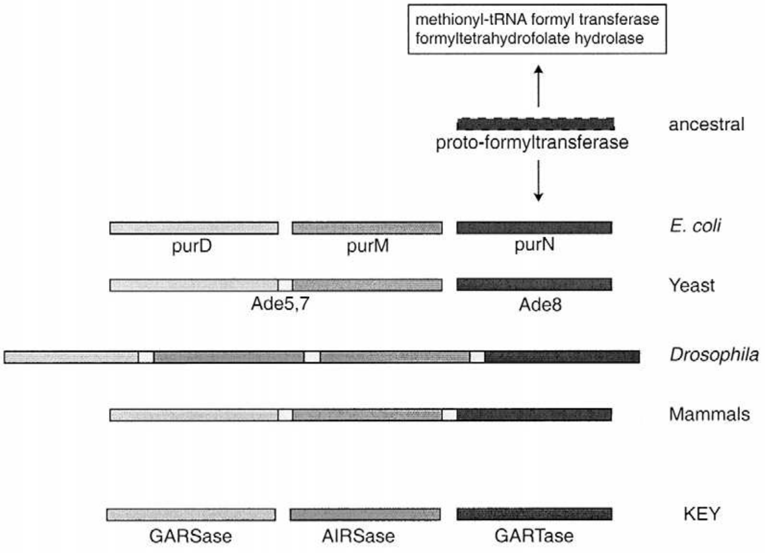


FIG. 2. Gene and domain organization of the glycinamide ribonucleotide synthetase (GARSase), aminoimidazole ribonucleotide synthetase (AIRSase), glycinamide ribonucleotide formyltransferase (GARTase), and related genes from different species. GARSase, AIRSase, and GARTase catalyze the second, fifth, and third steps in the *de novo* purine biosynthetic pathway. In mammals, chicken, and *Drosophila,* all enzymes are encoded by a single gene. The *Drosophila* gene differs by having an internal duplication of the AIRSase domain. In yeast, the GARSase and AIRSase are a single gene (Ade5,7) and GARTase is separate (Ade8). All three activities are encoded by separate genes in *E. coli.* Sequential, structural and functional similarities (see also Fig 4) suggest that fragments of the *E. coli* genes for GARTase, methionyl-tRNA formyltransferase, and formyltetrahydrofolate hydrolase have evolved from an ancestral formyltransferase and were incorporated into these genes by domain recruitment.

A. Multidomain Proteins

Multidomain proteins tend to occur more frequently in eukaryotes than in prokaryotes. Often the eukaryotic counterpart to a set of individual prokaryotic enzymes that catalyze successive reactions is a single, multidomain protein. The theoretical advantages proposed for such an arrangement include (1) a geometry for the direct transfer of substrates from one active site to another, in a process known as substrate channeling, in order to increase the overall flux of the pathway, (2) the protection of intermediates that may be unstable in aqueous environments or may be acted on inappropriately by other enzymes, (3) the facilitation of interactions between domains for purposes of allosteric regulatory functions, and (4) the establishment of a fixed stoichiometric ratio of the

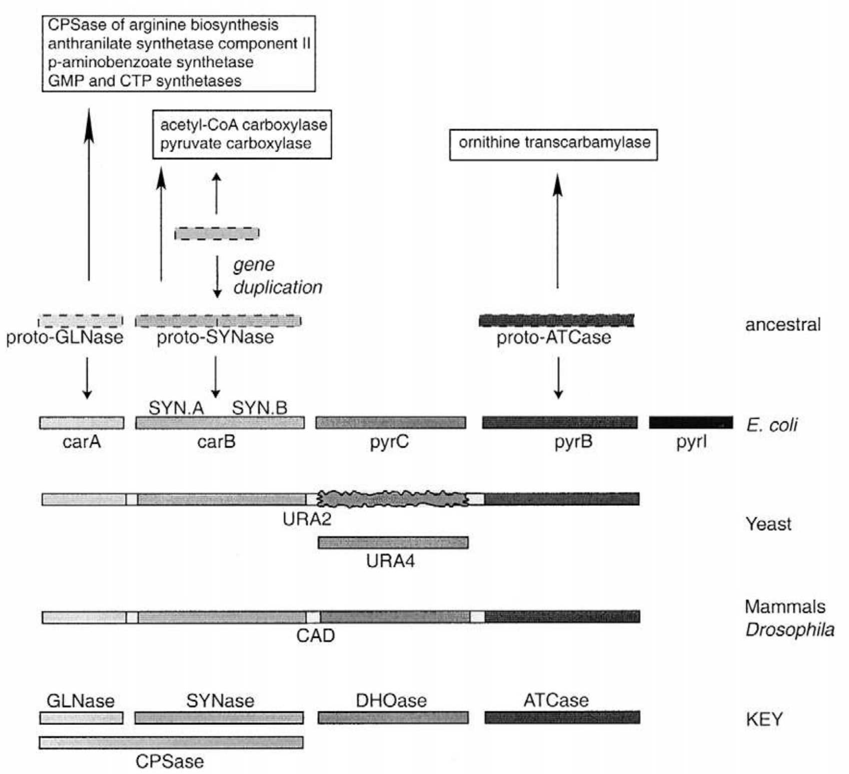


FIG. 3. Gene and domain organization of the carbamoyl-phosphate synthetase (CPSase), dihydroorotate dehydrogenase (DHOase), aspartate carbamoyltransferase (ATCase), and related genes from different species. CPSase, DHOase, and ATCase catalyze the first, third, and second steps in the *de novo* pyrimidine biosynthetic pathway. CPSase can be divided into domains with glutaminase (GLNase) and synthetase (SYNase) activities. In mammals and *Drosophila,* all activities are encoded by a single gene (CAD). The yeast URA2 gene codes for a similar length peptide, but the DHOase-like domain is inactive. In yeast, DHOase activity is encoded by the URA4 gene. All activities in *E. coli* are encoded by separate genes and include a regulatory submit for ATCase (pyrI). Sequential, structural, and functional similarities suggest that the CPSase and ATCase genes are evolutionarily related to other enzymes. GLNase is related to several enzymes that utilize cleavage of glutamine as the basis for catalysis. Acetyl-CoA carboxylase and pyruvate carboxylase contain domains with homology to the amino terminal half of SYNase, itself a product of gene duplication. This makes it likely that all three genes arose by domain recruitment from an ancestral proto-SYNase. Sequence similarities throughout the ornithine transcarbamylase and ATCase genes, both of which use carbamoyl phosphate as a substrate, suggest that they arose by duplication of an ancestral gene and divergent evolution.

enzymatic activities necessary for a sequential set of reactions (Srere, 1987; Hawkins and Lamb, 1995). Two examples of multidomain proteins are discussed: eukaryotic and prokaryotic enzymes in the *de novo* purine and pyrimidine biosynthetic pathways.

1. *Purine Biosynthesis*

Purines are synthesized *de novo* from phosphoribosyl pyrophosphate. The first ten enzymatic steps lead to the formation of inosine monophosphate, from which adenosine monophosphate and guanosine monophosphate are then synthesized in two steps. The second, third, and fifth steps in this pathway are catalyzed by glycinamide ribonucleotide synthetase (GARS), glycinamide ribonucleotide formyltransferase (GART), and aminoimidazole ribonucleotide synthetase (AIRS), respectively. In *Escherichia coli* (Zalkin and Nygaard, 1999) and *Bacillus subtilis* (Ebbole and Zalkin, 1987) these enzymatic activities lie on separate proteins, whereas in human, chicken (Aimi *et al.,* 1990) and mouse (Kan *et al.,* 1993), a single gene locus codes for a trifunctional protein with the domain order GARS-AIRS-GART (Fig. 2). Sequence and structural data clearly indicate that the individual genes from bacteria and the gene fragments corresponding to the functional domains of mammals and chicken have a common evolutionary origin. Examination of the organization of the corresponding genes in *Drosophila* (Henikoff *et al.,* 1986) and yeast (Henikoff, 1986) suggests a two-step evolutionary pathway for the fusion of these genes. Initially, as seen in yeast, genes for GARS and AIRS activities were fused, with GART activity remaining on a single gene. Subsequently the GARS-AIRS and GART genes fused. In *Drosophila,* tandem duplication of the AIRS region has led to a protein with two tandem AIRS domains. How the genes for the GARS-AIRS-GART multienzymatic protein became fused is an open question. A reasonable model has been proposed that employed intron-intron mediated rearrangement as the evolutionary mechanism, based largely on the presence of introns at domain boundaries (Davidson and Peterson, 1997).

2. *Pyrimidine Biosynthesis*

Carbamoyl-phosphate synthetase (CPSase), aspartate transcarbamoylase (ATCase), and dihydroorotase (DHOase) catalyze the first, second, and third steps in the pyrimidine *de novo* biosynthetic pathway, respectively. CPSase has two enzymatic functions that are found as separate units in bacteria: glutamine amidotransferase (GLNase) and synthetase (SYNase). GLNase transfers the amino group of glutamine to the catalytic site of SYNase, which in turn catalyzes the formation of carbamoyl phosphate in a complex reaction. GLNase- and SYNase-like domains are found in CPSase domains of higher species. CPSase, ATCase and CHOase are independent in bacteria but fused into multienzyme proteins in eukaryotes (CAD) in the order CPSase-DHOase-ATCase (Fig. 3). Unlike in the purine enzymes previously described, there is no intermediate level of gene fusions in species between bacteria and human to aid in the elucidation of the order of gene assembly. Interestingly, the yeast multifunctional enzyme URA2 contains an inactive DHOase-like domain between the CPSase and ATCase domains with an individual DHOase protein coded for on a separate operon (Souciet *et al.,* 1989). This observation is clearly suggestive of a gene duplication event.

Although channeling can be found in these multifunctional proteins, it seems that the creation of more sophisticated allosteric control was the driving force behind gene consolidation. Several lines of evidence indicate that URA2 channels carbamoyl phosphate from the catalytic site of CPSase to the catalytic site of ATCase (Belkaïd *et al.,* 1987; Hervé *et al.,* 1993) and partial channeling has been observed in mammalian CAD (Mally *et al.,* 1980). However, evidence for channeling has been observed in the hyperthermophilic archaeon *Pyrococcus abyssi* CPSase, ATCase, and DHOase enzymes, which are separate enzymes (Purcarea *et al.,* 1999), and in a truncated CAD with ATCase as a separate domain (Davidson *et al.,* 1993). Studies on independent domains of yeast and mammalian CAD and yeast-mammalian chimeric CAD indicate that catalytic activity resides entirely within the independent domains but that regulation requires interdomain interactions (Serre *et al.,* 1999). Thus, it seems that the evolutionary driving forces behind consolidation of these genes into a multifunctional enzyme were incorporating more sophisticated allosteric control and, presumably, coordinate expression.

B. Tandem Duplication

Duplication of genes or gene fragments in tandem is one method observed in nature to evolve new function. Its occurrence in so many proteins clearly supports its advantages: (1) increased stability, (2) new cooperative functions, (3) formation of a binding site in a cleft, (4) production of multiple binding sites in series resulting in more efficient or specific binding, and (5) growth of long repetitive structures in modular proteins (McLachlan, 1987). The two possible evolutionary outcomes for such duplications are domains that function in isolation and domains that interact. The former are likely to exist initially after duplication, with interactions, if evolutionarily favorable, developing concomitant with genetic changes in the two domains. Tandem duplication is also a proposed first step in a mechanism for circular permutation of proteins (Ponting and Russell, 1995).

1. *Pyrimidine Biosynthesis*

CPSase catalyzes the formation of carbamyl phosphate from glutamine, bicarbonate, and two equivalents of ATP. The biosynthesis involves four partial reactions. GLNase catalyzes the formation of ammonia from glutamine. The remaining three partial reactions are catalyzed by SYNase. Bicarbonate is activated by ATP to form carboxyphosphate, which reacts with ammonia to form carbamate. The ATP-dependent phosphorylation of carbamate results in the production of carbamyl phosphate.

Comparison of the N-terminal and C-terminal fragments of the *E. coli carB* gene, which codes for the large subunit of CPSase, suggests that it arose from tandem duplication of a smaller ancestral gene (Nyunoya and Lusty, 1983) (Fig. 3). The homology is especially strong between residues 1-400 and 553-933, which exhibit 39% identity and 64% homology with only four minor adjustments for insertions and deletions. Although these two domains are commonly referred to as CPS.A and CPS.B, they will be referred to as SYN.A and SYN.B, respectively, to avoid confusion with CPSase. Since the two ATP-dependent reactions of SYNase utilize similar substrates (bicarbonate and carbamate), one might assume that tandem duplication of an ancient SYNase half-domain occurred to catalyze these two reactions at different sites. However, the picture is more complicated than that.

Remarkably, the isolated GLNase-SYN.A domains and a GLNase-SYN.B fusion catalyze the entire series of reactions involved in glutamine-dependent carbamyl phosphate synthesis (Guy and Evans, 1996). Also, a truncated URA2 gene with the GLNase and SYN.A domains removed still codes for an active enzyme that retains the ability to channel carbamoyl phosphate and be allosterically regulated by UTP (Serre *et al.,* 1999). Furthermore, *Pyrococcus* archaebacteria posses a SYNase that is less than half the size of other SYNases (Purcarea *et al.,* 1996; Durbecq *et al.,* 1997). All this evidence suggests that the ability to synthesize carbamylphosphate predates tandem duplication.

This begs the question as to why the SYNase domain has evolved by tandem duplication, since both SYN.A and SYN.B seem to be functionally equivalent. However, evidence that the two ATP-dependent reactions occur at different sites in SYNase is very strong (Guy *et al.,* 1996), and it has been subsequently shown that the functional form of SYN.A and SYN.B domains in the absence of the other is a homodimer (Guy *et al.,* 1998). It appears that it is somehow advantageous to have separate, but proximal, domains for the ATP-dependent partial reactions. Although this can occur intermolecularly, fusing the two domains together ensures this geometry and stoichiometry. That both SYN.A and SYN.B can catalyze the two reactions is probably an evolutionary leftover from an ancestral single domain SYNase, like that of *Pyrococcus,* which catalyzed both reactions. This view of the function of SYN.A and SYN.B supports an evolutionary pathway by tandem duplication, followed by optimization by point mutation.

2. *Proteases*

The pepsin and chymotrypsin families of proteases are believed to have evolved by tandem duplication. Members of the pepsin family of proteases are composed of two large β-sheet domains, with their active sites in between composed of catalytic aspartate residues from each domain. Aside from structural similarities, the domains have sequence homology, particularly around the active site aspartate residues. This clearly suggests that the two domains arose by tandem duplication (Tang *et al.,* 1978). It seems that tandem duplication in pepsin allowed the creation of a binding site cleft. Members of the chymotrypsin family of proteases are composed of two homologous β-barrel domains packed together asymmetrically. The important residues for catalysis are His57, Asp102, and Ser195. The domain interface forms the active site with the first two catalytic residues on the N-terminal barrel and the last one on the C-terminal barrel. Superposing the α-carbon atoms of the two domains shows forty-six carbons fitting within a root-mean-square distance of 2.4 Å (McLachlan, 1979). This suggests that chymotrypsin evolved by tandem duplication, although no detectable sequence similarity remains.

3. *Modular Proteins*

Modular proteins are described as displaying a beads-on-a-string organization of domains or modules because the individual domains (beads) function independently but are connected via the peptide backbone (string). Such an organization has been found in many eukaryotic proteins such as fibronectin, collagen XII, factors involved in blood clotting and fibrinolysis, muscle associated proteins such as twitchin and titin, and cell surface receptors (Baron *et al.,* 1991; Doolittle and Bork, 1993; Hegyi and Bork, 1997). The demonstration that domain folding in prokaryotes is posttranslational and domain folding in eukaryotes is cotranslational and sequential may have been critical in the evolution of modular proteins by allowing tandem duplication events to create immediately foldable protein structures (Netzer and Hartl, 1997).

C. Domain Recruitment

Domain recruitment is a mechanism by which functional units from one protein are ''recruited'' by another protein. This mixing and matching of existing domains constitute an efficient method to evolve proteins. An apt explanation of this is a variation on the saying that ''a thousand monkeys typing at a thousand typewriters would eventually reproduce the works of Shakespeare.'' The monkeys would obviously work much more efficiently if, once they managed to type a coherent word, sentence, paragraph, or chapter, they could reproduce these words, sentences, paragraphs, and chapters with a single keystroke.

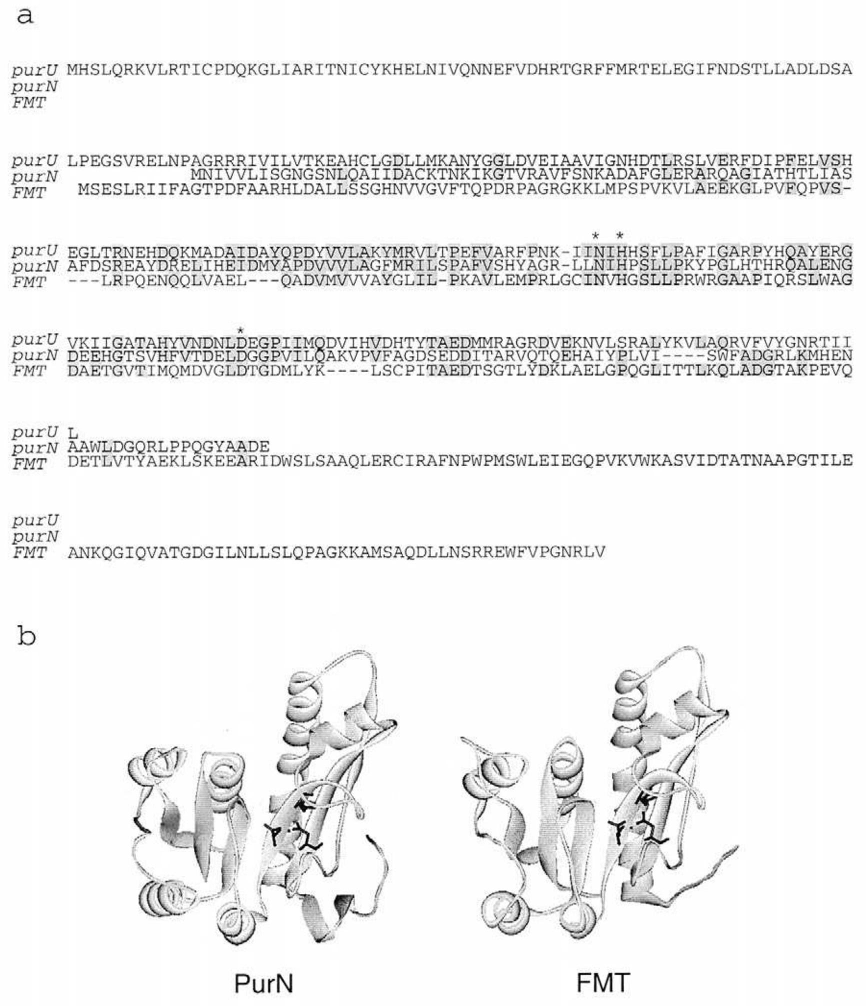
1. Purine Biosynthesis

The *E. coli* genes for glycinamide ribonucleotide transformylase (PurN), methionyl-tRNA formyltransferase (FMT), and formyltetrahydrofolate hydrolase (purU) catalyze the transfer of the formyl group from formyltetrahydrofolate to glycinamide ribonucleotide, methionyl-tRNA, and water, respectively. Sequence homology, conservation of catalytic residues, and structural similarities between PurN and FMT (see Fig. 4) suggest a common ancestor for the subdomain responsible for formyltetrahydrofolate binding and deformylation.

The N-terminal domain of PurN is structurally homologous to other N-terminal domains of GAR synthetase (PurD) and N^5-carboxyaminoimidazole ribonucleotide synthetase (PurK)—*E. coli* enzymes that are involved in purine biosynthesis (Wang *et al.,* 1998; Thoden *et al.,* 1999). This N-terminal domain is responsible for ribonucleotide binding and adopts a Rossman-fold (Rossmann *et al.,* 1974) that is common in nucleotide binding proteins. As a result of the structural similarity in the N-terminus of three enzymes in the *de novo* purine biosynthetic pathway, it is tempting to speculate that these enzymes have evolved by domain recruitment. However, the very low sequence homology makes this difficult to prove. If domain recruitment did occur, it must have occurred with an early ancestral nucleotide-binding motif for such a divergence in sequence to accumulate.

2. Pyrimidine Biosynthesis

ATCase and ornithine carbamoyltransferase (OTCase) catalyze analogous reactions. ATCase transfers the carbamoyl moiety from carbamoyl phosphate to aspartate, and OTCase transfers the carbamoyl moiety from carbamoyl phosphate to ornithine. They both share a common N-terminal functional domain, which binds carbamoyl phosphate. The C-terminal domains of these enzymes are structurally similar but have



```
purU MHSLQRKVLRTICPDQKGLIARITNICYKHELNIVQNNEFVDHRTGRFFMRTELEGIFNDSTLLADLDSA
purN
FMT

purU LPEGSVRELNPAGRRRIVILVTKEAHCLGDLLMKANYGGLDVEIAAVIGNHDTLRSLVERFDIPFELVSH
purN            MNIVVLISGNGSNLQAIIDACKTNKIKGTVRAVFSNKADAFGLERARQAGIATHTLIAS
FMT   MSESLRIIFAGTPDFAARHLDALLSSGHNVVGVFTQPDRPAGRGKKLMPSPVKVLAEEKGLPVFQPVS-

                                                       *  *
purU EGLTRNEHDQKMADAIDAYQPDYVVLAKYMRVLTPEFVARFPNK-IINIHHSFLPAFIGARPYHQAYERG
purN AFDSREAYDRELIHEIDMYAPDVVVLAGFMRILSPAFVSHYAGR-LLNIHPSLLPKYPGLHTHRQALENG
FMT  ---LRPQENQQLVAEL---QADVMVVVAYGLIL-PKAVLEMPRLGCINVHGSLLPRWRGAAPIQRSLWAG

                    *
purU VKIIGATAHYVNDNLDEGPIIMQDVIHVDHTYTAEDMMRAGRDVEKNVLSRALYKVLAQRVFVYGNRTII
purN DEEHGTSVHFVTDELDGGPVILQAKVPVFAGDSEDDITARVQTQEHAIYPLVI----SWFADGRLKMHEN
FMT  DAETGVTIMQMDVGLDTGDMLYK----LSCPITAEDTSGTLYDKLAELGPQGLITTLKQLADGTAKPEVQ

purU L
purN AAWLDGQRLPPQGYAADE
FMT  DETLVTYAEKLSKEEARIDWSLSAAQLERCIRAFNPWPMSWLEIEGQPVKVWKASVIDTATNAAPGTILE

purU
purN
FMT  ANKQGIQVATGDGILNLLSLQPAGKKAMSAQDLLNSRREWFVPGNRLV
```

FIG. 4. (a) Sequence alignment of the *E. coli* genes for glycinamide ribonucleotide formyltransferase (*purN*), methionyl-tRNA formyltransferase (*FMT*), and formyltetrahydrofolate hydrolase (*purU*). Identical amino residues are shaded in gray. The key active site residues of PurN are indicated by (*). The N-termini of all three enzymes do not align well, but a sequence alignment of the N-termini of *purN* and *FMT* based on a structural alignment (not shown here) has been proposed (Schmitt *et al.,* 1996). (b) Crystal structures of PurN (Almassy *et al.,* 1992) and the domain of FMT (Schmitt *et al.,* 1996) homologous to PurN. The C-terminal domain of FMT is not shown. Side chains of the key active site residues of PurN (Asn106, His108 and Asp144) and FMT (Asn108, His110 and Asp 146) are shown in black.

very divergent sequences. Based on multiple sequence alignments, the simplest explanation is that the two carbamoyltransferases arose not by domain recruitment but by duplication of an ancestral gene followed by divergent evolution (Labedan *et al.,* 1999) (Fig. 3).

In contrast, sequence relationships between CPSase and enzymes outside the pyrimidine pathway suggest domain recruitment as an evolutionary mechanism. Homology between the N-terminal half of SYNase and subdomains of acetyl-CoA carboxylase (Takai *et al.,* 1988) and pyruvate carboxylate (Lim *et al.,* 1988) suggest that these enzymes evolved by domain recruitment. Similarly, GLNase is sequentially and functionally related to domains in several enzymes that utilize cleavage of glutamine as the basis for catalysis. GLNase probably evolved from an ancestral glutaminase that was duplicated and inserted into other proteins by domain recruitment (Davidson *et al.,* 1993).

3. Insertion of Domains

Identification of homologous domains can be confounded by irregularities in the relationships between linear sequences and three-dimensional structures (Russell and Ponting, 1998). This arises primarily from the assumption that domains form distinct compact structures that are coded for by a single, uninterrupted stretch of DNA. Although end-to-end fusion is the predominant manner of linking two domains, the increasing number of exceptions to this rule suggests an evolutionary tolerance or even advantage to domain insertions. Although a protein created by a domain insertion would be expected to incur protein folding problems for entropic reasons, a number of advantages for insertions have been proposed (Russell, 1994). Fixing both ends of an inserted domain offers the advantages of a more compact structure to avoid proteolysis and a more rigid structure to fix the relative spatial orientation of the two domains in order to make functional combinations more viable. Finally, a possible reason why one third of all proteins have their N- and C-termini proximal (Thornton and Sibanda, 1983)—a proximity that is required for a domain to be inserted—could be to allow for domain insertion as an evolutionary mechanism.

Protein structure determinations have identified several examples of one domain inserted within another. One example is the *E. coli* DsbA protein, which catalyzes the formation of disulfide bonds in the periplasm. The enzyme consists of two domains: a thioredoxin-like domain that contains the active site, and an inserted helical domain similar to the C-terminal domain of thermolysins (Martin *et al.,* 1993). The inserted domain forms a cap over the active site, suggesting that it plays a role in binding to partially folded polypeptide chains before oxidation of cysteine residues. It has been proposed that DsbA's potent disulfide catalyzing ability results from the relative motions of the two domains, though recent evidence shows that the motions are independent of the redox state of the active site and thus may only be related to substrate binding (Guddat *et al.,* 1998). It seems likely that DsbA arose by the insertion of a domain into an ancestral thioredoxin-like domain in order to provide improved substrate binding, raising the catalytic power of the enzyme. Of course, an alternative explanation that must be considered is that thioredoxin is a result of the deletion of the helical domain from an ancestral DsbA protein (Russell, 1994). Other examples of domain insertion have been summarized elsewhere (Russell, 1994; Russell and Ponting, 1998).

D. Circular Permutations

A circularly permuted protein has its original N- and C-termini fused and new N- and C-termini created by a break elsewhere in the sequence. Circular permutation can be thought of as a form of domain swapping in which the C-terminal fragment of a gene has been moved to the beginning of that gene. Circular permutation is believed to result when tandem duplication of a gene is followed by the introduction of a new open reading frame within the first repeat and a new stop codon within the second repeat (see Fig. 1). This model is supported by the observation of tandem duplication in prosaposins and adenine-N^6 DNA methyltransferase, genes for which circular permutated variants are known (Ponting and Russell, 1995; Jeltsch, 1999).

The first *in vitro* construction of a circular permuted protein was carried out on bovine pancreatic trypsin inhibitor by chemical means (Goldenberg and Creighton, 1983). Since then, a number of circular permuted proteins have been constructed, primarily by genetic methods (for a review, see Heinemann and Hahn, 1995). These studies have shown that circular permuted proteins very often fold up into stable, active proteins. Comparisons of primary and tertiary structures within several protein families have led to the conclusion that circular permutation occurs in natural protein sequences. A recent review noted evidence of naturally occurring circular permutations in lectins, bacterial β-gluconases, aspartic proteases, glucosyltransferases, β-glucosidases, surface layer homology domains, transaldolases, and C2 domains (Lindqvist and Schneider, 1997). Another recent example is adenine-N^6 DNA methyltransferases (Jeltsch, 1999).

An interesting case is a family of plant aspartic proteases in which sequence alignments have predicted evolution by insertion of a saposin-like domain that itself has been circularly permuted. Circularly permuted variations of saposin-like domains have been termed ''swaposins'' and were first identified in plant aspartic proteases (Guruprasad *et al.,* 1994). The four α-helices of saposin have been permuted in swaposin such that the order in swaposin is helices 3-4-1-2. The cDNA sequence that codes for saposins suggests a mechanism for the evolution of swaposin (see Fig. 1) (Ponting and Russell, 1995). Saposin domains are formed by cleavage of prosaposin, which consists of four tandem repeats of saposin domains. This is suggestive of an evolutionary event in which swaposin was created by the creation of a new ORF whose N-terminus is helixes 3 and 4 followed by a linker and helices 1 and 2.

The evolutionary advantages of circular permutation are not clear from the current examples of naturally occurring circularly permuted proteins, as most have conserved function. However, examples of engineered circularly permuted proteins have been shown to result in higher stability (Wieligmann *et al.,* 1998), effect quaternary domain assembly in multi-domain proteins (Wright *et al.,* 1998), and increase antitumor activity in fusions of interleukin 4 and *Pseudomonas* exotoxin by changing of their topology (Kreitman *et al.,* 1995).

IV. Domain Swapping for Protein Engineering

The following sections describe examples of domain swapping performed to engineer proteins with desired function and properties. Studies are categorized under the intended goals of the protein engineering: engineering of allosteric regulation, creation of activators and inhibitors, improvement in stability or expression, modification of substrate specificity, improvement of catalytic efficiency, alteration of multimodular synthetases, improvement of therapeutic properties, creation of molecular biosensors, and creation of novel enzymes. Some of this work has also been explored in two recent reviews on hybrid enzymes (Nixon *et al.,* 1998; Beguin, 1999).

Bifunctional enzymes, generally end-to-end fusions of two domains or proteins in which the two proteins have distinct and independent functions, are mentioned only briefly in order to devote more space to domain-swapped enzymes that have a higher degree of interaction. Recent reviews that discuss the use of bifunctional enzymes in more detail include those on recombinant immunotoxins (Reiter and Pastan, 1998), signal sequences for cellular targeting (Martoglio and Dobberstein, 1998), gene fusions for phage display (Hoogenboom *et al.,* 1998; Rodi and Makowski, 1999) or bacterial display (Georgiou *et al.,* 1997), affinity purification (Nilsson *et al.,* 1997), improved recombinant protein production (Baneyx, 1999), *in vivo* detection of proteins (Tsien, 1998), chimeric transcription activators (Clackson, 1997), artificial systems for assembling enzymes (Beguin, 1999), and two- and three-hybrid systems (Drees, 1999). Similarly, coverage will be minimal of hybrids between interspecies homologs that were constructed for purposes of understanding differences in their properties.

A. Engineering of Allosteric Regulation

The engineering of new allosteric regulations by domain swapping or other methods has not been extensively explored, but two studies suggest its general feasibility. First, insertion of TEM β-lactamase into two different loops of the *E. coli* maltose binding protein (MalE) have been found to result in β-lactamase activity that is more stable to urea denaturation in the presence of maltose than in its absence (Betton *et al.,* 1997). Second, heterotetramers of L-lactate dehydrogenase (LDH), which naturally forms homotetramers, and an engineered form of LDH with the substrate specificity of malate dehydrogenase (MDH) exhibit allosteric properties (Fushinobu *et al.,* 1998). MDH activity in these LDH/MDH mixed tetramers was found to increase in the presence of oxamate, which was found to bind to LDH in the complex.

An enzymatic two-hybrid system, termed protein-fragment complementation assay (PCA) (Pelletier *et al.,* 1998) has potential for engineering of allosteric regulation. Developed to link protein interactions to enzymatic activity, PCA has elements of both domain swapping and regulation of enzyme activity. PCA consists of two designed fragments of murine dihydrofolate reductase (mDHFR), each of which is fused to a target domain (i.e., two domains that potentially interact). These two mDHFR domains were engineered by circular permutation followed by cleavage at a designed location. The two mDHFR domains without the target domains do not associate efficiently enough to produce DHFR activity. Association between the two target domains drives the association of the mDHFR fragments. Since mDHFR activity is essential for complementation of *E. coli* grown in the presence of the anti-folate drug trimethoprim, domains that interact can be selected by functional complementation. Allosteric regulation dependent on a small molecule can be created if the association between the two interacting domains is dependent on a small molecule. Such a case was demonstrated, though not for the purposes of regulating DHFR activity, with the FK506 binding protein (FKBP) and a domain of the FKBP-rapamycin binding protein (FRB) fused to the mDHFR fragments (Remy *et al.,* 1999). FKBP and FRB only associate in the presence of rapamycin. Hence, the mDHFR

PCA can be used for quantitative pharmocological analysis of protein–protein and protein–small molecule binding *in vivo.* However, if the mDHFR domains are swapped for complementary fragments of another enzyme, then perhaps a PCA with ligand-dependent interacting domains such as FKBP and FRB can be used to regulate the activity of this enzyme.

B. *Creation of Activators and Inhibitors*

In *Drosophila,* three proteins with epidermal growth factor-like domains modulate the function of the epidermal growth factor (EGF) receptor. Spitz (Spi) is a potent activator, Vein (Vn) is a moderate activator, and Argos (Aos) is an inhibitor. Aside from the EGF-like domains, these three proteins are structurally unrelated, suggesting that they evolved by EGF-like domains being fused to ancestral Spi, Vn, and Aos. Swapping of the EGF-like domains from Spi and Aos into Vn resulted in hybrids that behaved *in vitro* and *in vivo* like Spi and Aos (Schnepp *et al.,* 1998). This work demonstrates the feasibility of swapping homologous structures between proteins containing other domains that are structurally unrelated.

C. *Improvement in Stability or Expression*

1. *Pyrroloquinoline Quinone Glucose Dehydrogenase*

Pyrroloquinoline quinone glucose dehydrogenase (PQQGDH), pyrroloquinoline quinone ethanol dehydrogenase(PQQEDH), and pyrroloquinoline quinone methanol dehydrogenase (PQQMDH) use pyrroloquinoline quinone (PQQ) as a bound cofactor. These enzymes have potential for use in diagnostics and as biosensors for organic compounds, such as using PQQGDH as an enzyme sensor for glucose. Based on the crystal structure of PQQMDH (Ghosh *et al.,* 1995) and predicted structures of PQQGDH and PQQEDH (Cozier and Anthony, 1995a; Cozier *et al.,* 1995b), all are β-propeller proteins with eight W-motifs (propellers) of four β-sheets each. PQQGDH are found in a variety of bacteria and, despite their sequence homology, have a great deal of variety in PQQ binding stability, thermal stability, and substrate specificity (Yoshida *et al.,* 1999).

To delineate regions of the protein responsible for the observed differences and create PQQGDH enzymes with desired substrate activities and stability, Sode and co-workers have constructed a variety of domain-swapped hybrids between the *E. coli* and *Acinetobacter calcoaceticus* PQQGDHs. They identified the region between 32% and 59% of the

N-terminus as being responsible for EDTA tolerance (PQQ binding stability) in the *A. calcoaceticus* enzyme. Based on this data, a domain-swapped enzyme was constructed that exhibited *E. coli* enzyme specificity and *A. calcoaceticus* EDTA tolerance (Yoshida *et al.,* 1999). In addition, the C-terminal 3% of the *A. calcoaceticus* enzyme was found to confer thermal stability to the *E. coli* enzyme when swapped for the *E. coli* enzyme's C-terminal 3% (Sode *et al.,* 1995). Interestingly, the 32%–59% region also conferred thermal stability. A domain-swapped enzyme containing the 32%–59% and C-terminal 3% regions of *A. calcoaceticus* but elsewhere derived from *E. coli* had the highest thermal stability while maintaining an EDTA tolerance that was essentially that of *A. calcoaceticus* (Yoshida *et al.,* 1999).

This work utilizes a functional definition of domains rather than a topological one. Mapping the swapped regions onto the predicted structure of PQQGDH shows that the region between 32% and 59% correlates to the second, third, and fourth W-motifs of the eight W-motifs, and the C-terminal 3% correlates to part of the second-to-last β-sheet in the eight W-motif. It is important to note that although the engineered enzyme had the desired thermal stability, EDTA tolerance, and specificity, activity toward glucose had decreased to 6% of the level of the *E. coli* enzyme as judged by $V_{\max}/K_m$. This correlated with an increase in random structure as judged by circular dichroism. Large reductions in activity are likely to be common in domain-swapped enzymes with lower levels of homology. However, directed evolution strategies to improve activity, which have been particularly successful in improving the activity of poor enzymes (Arnold and Volkov, 1999), will likely be successful in fine-tuning domain-swapped enzymes.

2. *Other Examples*

The three-dimensional structure of human extracellular superoxide dismutase (EC-SOD) is unknown. Studies of structure–function relationships have been severely limited by its poor production in mammalian cell lines and failure to be expressed in prokaryotic and yeast systems. In contrast, extra- and intracellular Cu- and Zn-containing superoxide dismutases (CuZn-SOD) are expressed very well in *E. coli* and yeast. CuZn-SOD is homologous to a large interior fragment of EC-SOD, but lacks its extra N-terminal and C-terminal domains. Fusions of either the N-terminal domain of EC-SOD or both the N- and C-terminal domains of EC-SOD to CuZn-SOD resulted in a domain-swapped enzyme that expressed well and whose characteristics resemble EC-SOD (Stenlund and Tibell, 1999).

Domain swapping can also result in the acquisition of proteolytic stability. Insertion of TEM β-lactamase into loops of the *E. coli* maltose binding protein (MalE) results in fusions that are less susceptible to proteolysis than simple end-to-end fusion (Betton *et al.,* 1997). The stability of the β-lactamase activity in the insertion proteins to urea induced denaturation was greater in the presence of maltose than in its absence. In addition, fusion partners, particularly MalE, can greatly improve the solubility and expression of aggregation-prone proteins (Kapust and Waugh., 1999). Engineered circular permutation almost always results in a slight to significant decrease in stability, but a recent exception to the rule has been found. Circular permutation of eye lens βB2-crystallin resulted in a modest improvement of stability to urea denaturation, with a midpoint of transition of 2.1 M compared to 1.9 M for the wild-type crystallin (Wieligmann *et al.,* 1998). The increased stability was attributed to a more compact structure of the individual domains of the protein.

D. Modification of Substrate Specificity

1. Chimeric Restriction Endonucleases

One of the greatest successes in utilizing domain swapping for protein engineering has been the creation of chimeric restriction endonucleases by Chandrasegaran and co-workers (Fig. 5) (Chandrasegaran and Smith, 1999). Although most known restriction enzymes recognize a particular DNA sequence and then cleave the DNA within that sequence, type IIS restriction enzymes recognize a particular site and then cleave the DNA a certain number of bases away from that site, irrespective of the cleavage site's sequence. This property is suggestive of two domains: a DNA-binding domain responsible for recognition and specificity, and a non-specific endonuclease domain. The type IIS restriction enzyme *Fok*I has been shown by biochemical experiments (Li *et al.,* 1992; Li *et al.,* 1993) and confirmed by recently solved crystal structures (Wah *et al.,* 1997; Wah *et al.,* 1998) to display such a domain structure. This structural arrangement suggests that *Fok*I evolved by random fusion of a nonspecific endonuclease domain (F_N) and a DNA-binding domain.

A systematic method for constructing restriction enzymes that recognize a given desired sequence has obvious applications in molecular biology. More important, the construction of a restriction enzyme that specifically recognizes sequences of 16 to 18 bp, long enough to make it probable that such a sequence is unique in a genome, has important applications in gene therapy and the construction of transgenic animals.

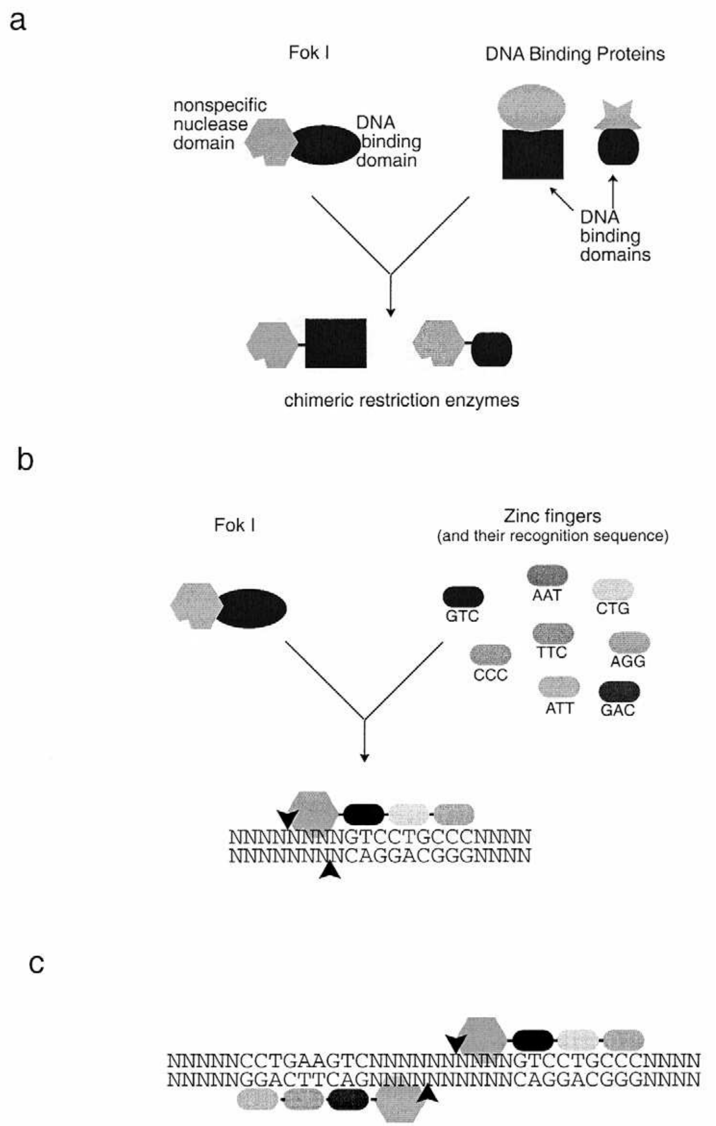

a
Fok I
DNA Binding Proteins
nonspecific nuclease domain
DNA binding domain
DNA binding domains
chimeric restriction enzymes
b
Fok I
Zinc fingers (and their recognition sequence)
GTC
AAT
CTG
CCC
TTC
AGG
ATT
GAC
NNNNNNNNNGTCCTGCCCNNNN
NNNNNNNNNCAGGACGGGNNNN
c
NNNNNCCTGAAGTCNNNNNNNNNNGTCCTGCCCNNNN
NNNNNGGACTTCAGNNNNNNNNNNCAGGACGGGNNNN

FIG. 5. Schematic of the construction and function of chimeric restriction endonucleases (Chandrasegaran and Smith, 1999). (a) The type IIS restriction enzyme *Fok*I has a nonspecific endonuclease domain (F_N) and a DNA-binding domain. Swapping *Fok*I's DNA binding-domain for that of another DNA binding-domain results in a chimeric restriction enzyme with new specificity. The flexibility resulting from less intimate contact between the F_N domain and its new DNA-binding domain translates into DNA cleavage at several locations near the binding site. (b) The modular nature and tunable specificity of zinc

Chandrasegaran and co-workers have made fine progress toward this goal by creating chimeric restriction enzymes using *Fok*I in which the natural DNA-binding domain has been swapped for a variety of DNA-binding domains, particularly zinc fingers (Chandrasegaran and Smith, 1999).

Functional chimeras have been constructed between F_N and the *Drosophila* Ubx homeodomain (Kim and Chandrasegaran, 1994), the DNA-binding domain of the transcription factor Sp1 (Huang *et al.,* 1996), yeast Gal4 (Kim *et al.,* 1998), and zinc-finger DNA-binding motifs (Kim *et al.,* 1996; Kim *et al.,* 1997). Zinc-finger motifs hold the most promise due to their modular nature (Fig. 5b). Crystal structures of zinc finger bound to DNA indicate that each zinc finger binds specifically to three bases (Pavletich and Pabo, 1991). This has stimulated efforts for creating a library of zinc fingers specific for each possible codon (Greisman *et al.,* 1997; Isalan *et al.,* 1997). Although fused zinc fingers do not bind totally independently, the linkage of several zinc fingers is believed to be an excellent framework to build DNA-binding proteins specific for sites up to 18 bp in length (Chandrasegaran and Smith, 1999). Sequence-specific cytosine methyltransferases were created by fusing zinc fingers and a CpG-specific DNA methyltransferase (Xu and Bestor, 1997).

A detailed study of a fusion of F_N and the designed zinc finger ΔQNK (a fusion of three zinc fingers) has demonstrated that the fusion does not significantly change the sequence specificity or binding affinity of the zinc fingers (Smith *et al.,* 1999). The sequence specificity of the F_N-ΔQNK chimeric endonuclease is a reflection of that of the ΔQNK. F_N-ΔQNK had no activity on some 9-bp sites with only one base difference from its target site. However, other sites with one base change were cleaved 16% as efficiently as the target site, as judged by the initial rate of cleavage. All chimeric F_N restriction enzymes described to date show nonspecific nucleotide activity, particularly at high $MgCl_2$ concentration and have multiple cleavage products (Kim *et al.,* 1998). This presumably results from the lack of interaction between the two domains in the chimera to ''lock'' the F_N domain in a fixed orientation relative to the DNA-binding domain: an interaction that is present in the natural *Fok*I enzyme. Constructing F_N-fusions with six zinc fingers in order to recog-

fingers potentially enables the construction of restriction enzymes for any desired sequence by the fusion of multiple zinc fingers together in a specific order. An F_N-three zinc finger chimeric restriction enzyme recognizing a nine-bp sequence is shown. (c) F_N-three zinc finger chimeric restriction enzymes that nick DNA but do not make double stranded breaks can be used in pairs to specifically recognize a noncontinuous 18-bp sequence.

nize a 18-bp site may achieve very tight binding, but at the expense of specificity. Interestingly, a zinc finger-F_N fusion with only DNA-nicking activity would allow circumvention of this problem. The use of two chimeras with nicking activity and different specificities potentially allows the specific cleavage within a noncontinuous 18-bp site (Fig. 5c) (Chandrasegaran and Smith, 1999).

2. *Proteases*

Members of the chymotrypsin family of serine proteases are composed of two homologous β-barrels that are believed to have evolved from an ancient gene duplication event (McLachlan, 1979). Numerous crystal structures of chymotrypsin-like proteases have shown that the two β-barrels stack together at roughly 90° to each other and are structurally conserved. The catalytic residues His57, Asp102, and Ser195 are at the interface of the two barrels, with His57 and Asp102 being from one domain and Ser195 being from the other. Variations between members of the chymotrypsin family primarily manifest themselves in surface loops near the active site that determine substrate specificity (Perona and Craik, 1997). Sites important for binding peptide substrates are distributed between the two domains.

Trypsin and factor Xa (fXa) are two members of the chymotrypsin family that have 38% sequence identity on the amino acid level and have distinguishable substrate specificities. Recently, the N-terminal β-barrel of fXa and the C-terminal β-barrel of trypsin were fused at a rationally designed site in the linker region between the two domains in order to create a hybrid fXa-trypsin protease (Hopfner *et al.,* 1998). The fXa-trypsin hybrid was highly active and more active than either parent on three of the ten substrates assayed, as determined by k_{cat}/K_m. For most substrates, the activity of fXa-trypsin was an admixture of the two parents, probably because trypsin had higher activity than fXa for all the substrates tested.

However, some of the most important results that impinge on domain swapping derive from the crystal structure of fXa-trypsin, which was determined to 2.15 Å. The general structural features of the chymotrypsin family are conserved between the respective parent molecules, including the core elements and the residues of the active site. Surface elements, particularly loops, exhibited much greater variability. Most important is the fact that although approximately 30% of the residues at the interface of the two domains differ between fXa and trypsin—presenting the possibility of unfavorable interactions that negatively effect folding and activity—these residues generally retained their structure with minor adjustments to accommodate mismatches. That plasticity apparently did not grossly affect stability or activity bodes well for a strategy of domain swapping that exchanges homologous structures irrespective of sequence homology. These results suggest that such exchanged structures are flexible enough to accommodate mismatches while generally retaining their fold. This possibly illustrates how large changes in sequence space can occur in nature, enabling the escape from local energy minimum and furthering the evolution of novel function.

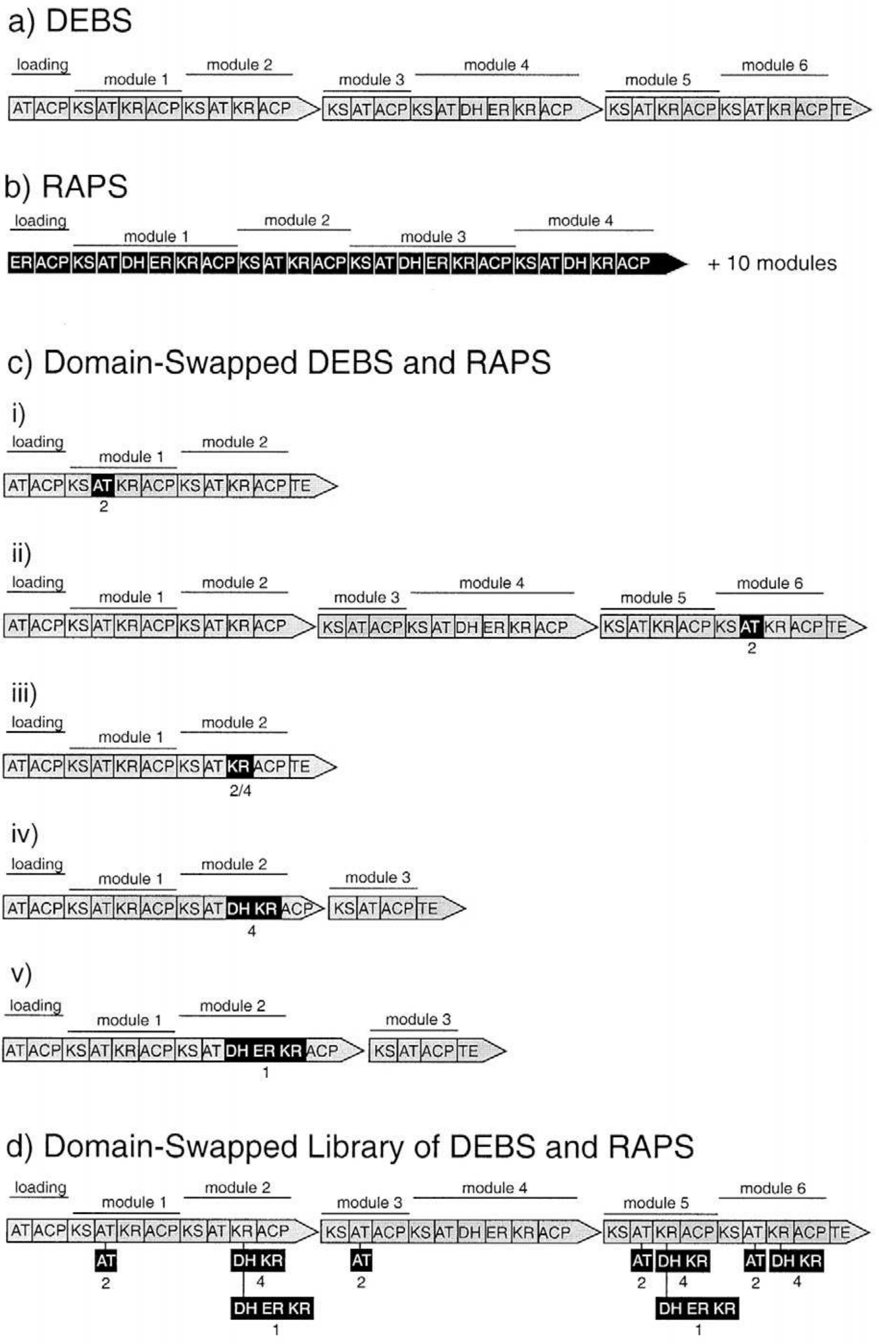
E. Improvement of Catalytic Efficiency

Studies on in-frame fusions between enzymes have demonstrated several benefits, including improved overall activity, pre-steady-state lag reduction, and sequestering of intermediates, presumably through the advent of substrate channeling (Bulow, 1998). Hybrids of β-galactosidase and galactose dehydrogenase (Ljungcrantz *et al.,* 1989), galactose dehydrogenase and luciferase (Lindbladh *et al.,* 1992), citrate synthetase and malate dehydrogenase (Lindbladh *et al.,* 1994), and two unique β-glucanases (Olsen *et al.,* 1996) have exhibited evidence of substrate channeling. All these bifunctional enzymes were constructed by end-to-end fusion of their parental genes with an intervening linker sequence. Presumably, this could be improved on by better organization of the active sites through variation of the linker length or through a change in the topology from end-to-end fusion to insertion of one enzyme into another. As the active sites are brought closer together, specific point mutations, perhaps identified through random mutagenesis, could be useful in developing interactions between the two domains.

F. Alteration of Multimodular Synthetases

1. Polyketide Synthases

Polyketides are made by the sequential activity of domains of large, multifunctional enzymes called polyketide synthases (PKSs) (Fig. 6a and b). Polyketides are formed by the condensation and modification of acyl units derived from acyl-CoA precursors. Domains are organized in modules and each module carries out the series of steps necessary for one cycle of polyketide chain elongation. A single protein can have more than one module, and several different proteins together can make up a PKS. The number of modules determines the size of the polyketide. A growing polyketide chain is tethered to the enzyme as a thiol ester and moves sequentially from the N- to the C-terminus of a module, lengthened by two carbon units per module. The first module in a PKS


a) DEBS
loading
module 1
module 2
module 3
module 4
module 5
module 6
AT ACP KS AT KR ACP KS AT KR ACP
KS AT ACP KS AT DH ER KR ACP
KS AT KR ACP KS AT KR ACP TE
b) RAPS
loading
module 1
module 2
module 3
module 4
ER ACP KS AT DH ER KR ACP KS AT KR ACP KS AT DH ER KR ACP KS AT DH KR ACP
+ 10 modules
c) Domain-Swapped DEBS and RAPS
i)
loading
module 1
module 2
AT ACP KS AT KR ACP KS AT KR ACP TE
2
ii)
loading
module 1
module 2
module 3
module 4
module 5
module 6
AT ACP KS AT KR ACP KS AT KR ACP
KS AT ACP KS AT DH ER KR ACP
KS AT KR ACP KS AT KR ACP TE
2
iii)
loading
module 1
module 2
AT ACP KS AT KR ACP KS AT KR ACP TE
2/4
iv)
loading
module 1
module 2
module 3
AT ACP KS AT KR ACP KS AT DH KR ACP
KS AT ACP TE
4
v)
loading
module 1
module 2
module 3
AT ACP KS AT KR ACP KS AT DH ER KR ACP
KS AT ACP TE
1
d) Domain-Swapped Library of DEBS and RAPS
loading
module 1
module 2
module 3
module 4
module 5
module 6
AT ACP KS AT KR ACP KS AT KR ACP
KS AT ACP KS AT DH ER KR ACP
KS AT KR ACP KS AT KR ACP TE
AT
2
DH KR
4
DH ER KR
1
AT
2
AT
2
DH KR
4
DH ER KR
1
AT
2
DH KR
4

has a loading domain that recognizes a particular acyl-CoA to start the chain. The mimimal module consists of ketosynthase (KS), acyltransferase (AT), and acyl carrier protein (ACP) domains. The specificity of the AT domain determines the α-alkyl moiety of the two-carbon unit. Each module can have between zero and three domains that are responsible for modifying the β-carbon of the polyketide chain. Those modules with modification domains can have a ketoreductase (KR), a KR and a dehydratase (DH), or a KR, a DH, and an enoylreductase (ER) domain. In the case when all three are present, the KR reduces the ketone to an alcohol, the DH dehydrates the alcohol to a double bond, and the ER converts the double bond to a saturate single bond.

Numerous studies have shown that the module nature of PKSs is very amenable for engineering new PKSs by domain swapping. Several recent reviews offer a detailed description of this work (Carreras and Santi, 1998; Hutchinson, 1999; Keating and Walsh, 1999). Much of the work has been with two PKSs: 6-deoxyerythronolide B synthase (DEBS; Fig. 6a) and rapamycin PKS (RAPS; Fig. 6b). The chain initiation is one step in the synthesis that can be altered by swapping loading domains. This was first demonstrated by replacing the loading domain of the platenolide PKS with that of the tylactone PKS, switching the specificity from acetate to propionate (Kuhstoss *et al.,* 1996). When the loading domain of DEBS was replaced with one from avermectin PKS, polyketides were obtained with side chains derived from a broader distribution of starter units, consistant with the avermectin PKS-loading unit's broader specificity for branched-chain carboxylic acids (Marsden *et al.,* 1998).

←

FIG. 6. Schematic of engineered polyketide synthases. Polyketide synthases (PKS) are made up of acyltransferase (AT), ketosynthase (KS), acyl carrier protein (ACP), ketoreductase (KR), dehydratase (DH), enoylreductase (ER), and chain termination (TE) domains that are grouped into modules. A single gene (indicated by the pointed wedges) can encode for more than one module. (a) The PKS 6-deoxyerythronolide B synthase (DEBS) has seven modules, including a loading module encoded by three genes (Cortes *et al.,* 1990; Caffrey *et al.,* 1992). (b) Rapamycin PKS (RAPS) has fifteen modules including a loading module encoded by three genes (Schwecke *et al.,* 1995). Only the first gene of RAPS is shown. (c) Many hybrid PKS have been engineered by replacing a domain from DEBS with domain(s) from RAPS. The number directly below the inserted domain(s) indicates the RAPS module from which the domain came. In (i), the AT domain from RAPS module 2 has replaced the AT domain of DEBS module 1 (Oliynyk *et al.,* 1996). Other functional hybrids include (ii) (Liu *et al.,* 1997), (iii) (Kao *et al.,* 1998), (iv) (McDaniel *et al.,* 1997) and (v) (Kao *et al.,* 1997). (d) A combinatorial library using three separate plasmids for the three genes of DEBS and the indicated array of domain swaps has been constructed (Xue *et al.,* 1999).

The AT domain is one logical place to alter PKS since it determines the α-alkyl moiety of the two-carbon unit that is added. The first demonstration of the feasibility of swapping AT domains came from the successful replacement of the AT domain of DEBS module 1 with the AT domain from RAPS module 2 that resulted in a polyketide with the intended substitution (Fig. 6c,i) (Oliynyk *et al.,* 1996). Replacements of the AT domain from DEBS module 6 with the AT domain from RAPS module 2 (Fig. 6c,ii) (Liu *et al.,* 1997), and AT domains of DEBS modules 1 and 2 with three different heterologous AT domains (Ruan *et al.,* 1997) resulted in novel polyketides being synthesized.

β-carbon-modifying domains (KR, DH, and ER) can also be swapped to produce novel polyketides. This can result in polyketides with stereochemical changes, such as changing an alcohol configuration from S to R by swapping the KR domain of DEBS module 2 for that of RAPS module 2 or 4 (Fig. 6c,iii) (Kao *et al.,* 1998). Modules can also be extended in length by, in effect, adding DH or DH and ER domains to modules that have only a KR domain such as DEBS module 2. Replacement with a KR and DH pair (Fig. 6c,iv) or KR, DH, and ER trio (Fig. 6c,v) from RAPS resulted in the expected products and thereby increased the functionality of the DEBS (McDaniel *et al.,* 1997; Kao *et al.,* 1997).

Recent publications in the field of engineered modular PKSs demonstrate the combinatorial potential of domain swapping to create novel PKSs to synthesize novel polyketides. Systematic replacement of AT and KR domains in modules 2, 5, and 6 of DEBS with AT and (KR or KR/DH or KR/DH/ER) domains from RAPS, respectively, resulted in a library of PKSs capable of synthesizing more than fifty different polyketides (McDaniel *et al.,* 1999). An approach for generating much larger libraries takes advantage of the fact that DEBS is encoded by three distinct genes whose proteins interact to form the PKS. Variants of individual genes are cloned into separate compatible plasmids, all of which are transformed into the same *Streptomysces lividans* strain. In this way, once the original set of variant genes are constructed, a library of all combinations of these variant proteins can be obtained in a relatively short amount of time by transforming cells with all three plasmid types (Fig. 6d). To demonstrate this system, fourteen plasmids were constructed and combinatorialized. Sixty-four cells received all three plasmids, of which forty-six produced detectable levels of forty-three different polyketides (Xue *et al.,* 1999). A combinatorial method of generating the original variants would significantly expand this combinatorial approach.

2. *Peptide Synthetases*

Nonribosomal peptide synthetases have a modular structure and carry out the syntheses of peptides about two to fifteen amino acids in length through the sequential action of several semi-independent domains, each of which carries out a specific reaction on the growing polypeptide chain. Distinct domains have been shown to be involved in amino acid recognition and adenylation (A-domain), covalent linkage of the substrate to the cofactor 4′-phosphopantetheine (T-domain), condensation of two activated petidyl or aminoacyl moieties (C-domain), and release of the product by cyclization or hydrolysis (Te-domain) (Mootz and Marahiel, 1999). The nature, number, and order of these domains determine the structure of the product.

Although not as extensively explored as recombining polyketide synthases has been, redesign of peptide synthetases by recombining their modules has been successful in a handful of studies (Stachelhaus *et al.,* 1995; de Ferra *et al.,* 1997; Schneider *et al.,* 1998) and was the subject of a recent review (Mootz and Marahiel, 1999). The first example of successful engineering of an active peptide synthetase to produce a new product was engineered by swapping A-domain and T-domains (AT-domain) from different species for the AT-domain of the seventh module of the seven module SrfA complex in *Bacillus subtilis* (Stachelhaus *et al.,* 1995). Phe-, Orn- and Leu-AT-domains from *Bacillus brevis* and Cys- and Val- AT-domains from *Penicillium chrysogenum* were swapped for the Leu-AT-domain. These five domain-swapped enzymes encoded peptide synthetases with the desired amino-acid specificities and produced a surfactin (containing seven amino acid moieties including two D-amino acids) with the desired amino acid substitutions, albeit at lower levels. Subsequent experiments in which AT-domains were swapped for the second and fifth position of the SrfA complex also resulted in an altered, but not always the expected, product (Schneider *et al.,* 1998). Swapping the Te-domain (responsible for release of the product from the seventh module) for the fourth and the fifth modules resulted in the expected shortened peptide products (de Ferra *et al.,* 1997). Fusion points chosen at estimated domain boundaries (Stachelhaus *et al.,* 1995; Schneider *et al.,* 1998), within highly conserved motifs (Ritsema *et al.,* 1998), and at fortuitously located restriction sites (de Ferra *et al.,* 1997) have been successful for creating peptide synthetases with some activity. For creating improved domain-swapped peptide synthetase, identifying the optimal choice of fusion points is likely to be crucial (Mootz and Marahiel, 1999) and may be well suited for the combinatorial methods described later.

G. *Improvement of Therapeutic Properties*

The spatial orientation of fused domains can affect important properties of the fusions. The use of circularly permuted domains overcomes the limited number of spatial orientations available by fusion at natural N- and C- termini. For example, fusion proteins of interleukin-4 (IL-4) and a *Pseudomonas* exotoxin (PE) have antitumor activity. However, the fusion point of the two domains is near the site of IL-4 that binds its receptor, limiting its effectiveness. However, circular permutation of Il-4 at residues 37 and 38 and fusion of PE to the new C-terminus resulted in a chimera that bound its receptor with ten-fold higher affinity (Kreitman *et al.,* 1994). It possessed improved, specific antitumor activity in a variety of different types of tumor cell cultures and increased efficacy in mice (Kreitman *et al.,* 1995). Similarly, granulocyte stimulating factor (GCSF) was circularly permuted at several locations and fused to interleukin-3 (IL-3) in an attempt to make a set of spatially permuted GCSF-IL-3 fusions (called myelopoietins (MPOs)) (McWherter *et al.,* 1999; Feng *et al.,* 1999). MPOs display enhanced activity to promote cell growth and maturation of hematopoietic cells. The activity of the circularly permuted GCSF ranged from 10% to greater than 100% of that of wild-type GCSF. Some of the fusions had altered ratios of GCSF to Il-3 activity, expanding the range of MPOs with therapeutic potential.

H. Creation of Molecular Biosensors

The green fluorescent protein (GFP) holds much potential for generating genetically encoded indicators for biochemical and physiological signals (Tsien, 1998). One possible arrangement to generate a signal is to fuse two variants of GFP, a donor and an acceptor with overlapping absorption and emission spectra, to the same or to interacting proteins. In this way, fluorescence resonance energy transfer (FRET) changes can be linked to molecular events that alter the distance (d) between the two GFPs. For example, variants of GFP have been fused to both ends of a calmodulin-binding domain (Romoser *et al.,* 1997) or to both ends of a calmodulin-(calmodulin-binding peptide M13) fusion (Miyawaki *et al.,* 1997) to generate a molecular sensor for Ca^{2+}. This domain undergoes large conformational changes on binding Ca^{2+}, effectively changing the distance between the two GFP variants such that a FRET signal can be linked to Ca^{2+} levels *in vitro* or *in vivo*.

Because FRET requires the two domains to be very close, proteins with their N- and C-termini distal may have difficulty in generating a FRET signal when the GFP variants are fused at both ends, since FRET efficiency decreases by a factor of d^6. Using such a system to monitor the interaction of two proteins, where each of the two variants of GFP is fused to different proteins, will have similar problems (Fig. 7a and b). Moving beyond a simple end-to-end fusion scheme can possibly bring

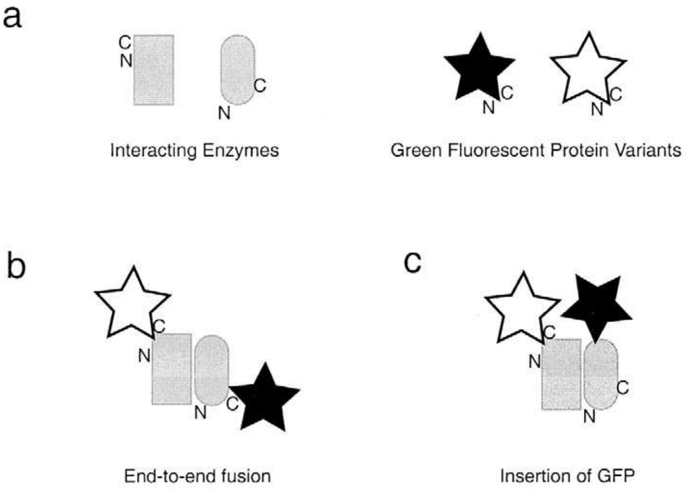


FIG. 7. Topological modifications to increase a FRET signal between two variants of the green fluorescent protein (GFP) fused to interacting proteins. (a) Schematic representation of interacting proteins and GFP variants with the location of the N- and C-termini indicated. (b) The GFP variants fused to the ends of the interacting proteins are too far apart to generate a FRET signal. (c) Insertion of one of the GFP variants into an interacting protein brings the two GFP variants close enough for FRET to be detected.

the two GFPs closer to allow the observance of FRET (Fig. 7c). To this end, random insertion of GFP into the cAMP-dependent protein kinase regulatory subunit from *Dictostelium discoideum* has been performed to increase the probability of observing FRET between the regulatory and catalytic subunits of this enzyme (Biondi *et al.,* 1998). However, this strategy did not result in any GFP insertion proteins with high fluorescence that maintained high affinity for cAMP. The binding of cAMP was still significant in many cases and 20% of the insertion proteins still interacted with the catalytic unit *in vitro.*

Even greater topological changes in GFP have been performed by random circular permutation (Graf and Schachman, 1996) in order to create new N- and C- termini for end-to-end fusion with other genes (Baird *et al.,* 1999). Ten nontrivial fluorescent circular permutations of GFP were found that had altered pK_a values and orientation of the chromophore with respective to its N- and C-termini. The systematic identification of sites for circular permutation in GFP also identifies plausible sites for insertions of other proteins into GFP. This work speaks strongly about the potential of random circular permutation for protein engineering and suggests a variety of protein topologies as targets for protein engineering (Baird *et al.,* 1999).

β-Lactamase has been inserted into a rationally chosen loop of GFP (a loop that, incidentally, was found to be a permissive site for circular permutation in the study described in the previous paragraph) in order to create a fusion protein whose fluorescence changes in the presence of β-lactamase-inhibitory protein (Doi and Yanagawa, 1999). Although in the initial construct no fluorescence change was observed on addition of the inhibitory protein, two rounds of directed evolution resulted in two mutations in the β-lactamase gene that conferred an inhibitory-protein concentration dependence on the fluorescence of the fusion. The mutated fusion had approximately 70% greater emission intensity in the presence of a tenfold molar excess of the inhibitory protein than in its absence. This is an excellent example of using domain swapping (by insertion) to move large distances in sequence space, followed by error-prone PCR to nudge the construct into a functional area of sequence space very nearby.

I. Creation of Novel Enzymes

Although all the examples of protein engineering previously discussed can be considered novel in some respects, the creation of an entirely novel enzyme by domain swapping has not been demonstrated. The phrase ‘‘entirely novel enzyme’’ is defined as an enzyme that displays catalytic activity that is neither present in nor meaningfully related to either parent. A theoretical example of creating an entirely novel enzyme by domain swapping would be to take a domain from protein A that is responsible for binding substrate 1 and fuse it to a domain from enzyme B to create a hybrid that could perform enzyme B’s chemistry on substrate 1. The natural substrates for enzyme B would bear little resemblance to substrate 1, and protein A would not have catalytic activity similar to enzyme B. Examples of engineered entirely novel enzymes created by methods other than domain swapping include the creation of catalytic antibodies (Smithrud and Benkovic, 1997) and the introduction of a protease catalytic triad into an *E. coli* cyclophilin to create a novel endopeptidase (Quemeneur *et al.,* 1998).

A step toward using domain swapping to create an entirely novel enzyme was taken by successfully creating a domain-swapped enzyme with glycinamide ribonucleotide (GAR) formyltransferase by fusing a ribonucleotide binding domain from PurN and a N^{10}-tetrahydrofolate cofactor hydrolase domain from PurU (Nixon *et al.,* 1997). Both PurN and PurU are formyltransferase enzymes (see Fig. 4). PurN transfers the formyl group of N^{10}-tetrahydrofolate to GAR to synthesize formyl-GAR and PurU transfers the formyl group to water to synthesize formate. Crystallographic, biochemical, and sequence comparisons were used to design several PurN–PurU hybrids. A PurN–PurU hybrid with GAR transformylase activity was characterized *in vitro* and found to have activity $\sim 10^4$ lower than PurN and to favor hydrolysis over transfer to GAR by a ratio of 40 : 1. Nevertheless, swapping a GAR binding domain into PurU and creating an active GAR transformylase is proof of principle that domain swapping to create novel activities is attainable. However, this work technically falls short of creating an entirely novel enzyme. This can be understood by instead viewing the active hybrid as derived from a swap of a formyltransferase domain of PurN for a formyltransferase domain of PurU.

V. Methodologies

Non-combinatorial methods for the engineering of domain-swapped enzymes have primarily been employed in the rational design of proteins. These methods achieve the fusion of one or more fragments at precise positions that are predicted through rational means to achieve a desired function. The probability of success for such an approach is directly proportional to the completeness of our understanding of protein folding, structure, function, and enzymatic mechanism in general and of the particular protein or protein fragments to be fused. Given that our knowledge is not complete, non-combinatorial methods are most likely to be successful in those areas in which structure and function of the fragments are not coupled and where desired function is objectively less complicated, as, for example, in the creation of bifunctional proteins for cellular targeting, phage or cell surface display, affinity purification, and increased expression/stability. Although combinatorial methods for optimizing linker length can be useful and have been employed, combinatorial methods seem more suited for achieving domain swapped enzymes in which the fragments are functionally coupled, such as in altering catalytic activity, modifying substrate specificity, engineering allosteric regulation and the creation of entirely novel enzymes.

A. Non-Combinatorial

For rationally constructing domain-swapped proteins, one consideration is whether to include a linker region between the two domains. In general, hybrids with two autonomously functioning domains require significant distance between the two domains so that they can fold and achieve a functional tertiary structure without interference from each other. The most common method for creating such domain-swapped enzymes is to clone the desired genes or gene fragments between properly arranged restriction sites such that the two fragments are separated by a piece of DNA that codes for a sequence of amino acids that can serve as a flexible linker. A consequence of this is that the DNA of the restriction sites will code for amino acids in the protein. However, appropriate choice of restriction sites that code for acceptable amino acids (e.g., *Bam*HI that codes for Gly-Ser, which are common linker amino acids) can go a long way to avoiding potential problems.

When the domains require more intimate contact to achieve the desired function, the presence of a linker and extra residues coded for by amino acids is not desirable. Although domain-swapped enzymes can be created by taking advantage of fortuitously located restriction sites or sites artificially introduced by silent mutagenesis, the simplest method for creating ‘‘seamless’’ fusions is by overlap extension (Horton *et al.,* 1989). Overlap extension can make end-to-end fusions in a single cloning step, but requires more than one cloning step to perform swapping of an internal domain. A variation of overlap extension for swapping internal domains in one cloning step has been described (Grandori *et al.,* 1997). Although other methods for making seamless fusions exist, such as vectors designed to allow the construction of precise fusions by creating unidirectional deletions between the two sequences to be fused (Kim *et al.,* 1991), they have not gained widespread use.

B. Combinatorial

The development of combinatorial processes for protein engineering that mirror natural evolution is a rational approach to achieving desired protein function. Homologous recombination is one method to create hybrids between highly homologous enzymes *in vivo* (Schneider *et al.,* 1981). *In vitro* methods of gene recombination, including DNA shuffling (Stemmer, 1994), staggered extension process (Zhao *et al.,* 1998), random priming recombination (Shao *et al.,* 1998), DNA reassembly by interrupting synthesis (Short, 1999), and restriction enzyme based shuffling (Kikuchi *et al.,* 1999), all differ in methods used to generate fragments to reassemble by sexual PCR. These methods depend on a stochastic process that reassembles genes based on homology and hence cannot combinatorialize genes with low DNA homology, cannot avoid biases due to homology, and cannot control maintenance of discrete domains or subdomains throughout the process. However, these methods clearly have utility in domain swapping, and are discussed in detail elsewhere in this volume.

1. *Random Insertions*

The observation of insertions in naturally occurring proteins suggests that such a route can be viable to construct proteins with desired properties and functions. Furthermore, studies on insertions of DNA coding for less than five amino acids (Zebala and Barany, 1991; Hallet *et al.,* 1997), up to sixteen amino acids (Starzyk *et al.,* 1989; Ladant *et al.,* 1992), a randomized 120 amino acids library (Doi *et al.,* 1997), and even entire proteins (Betton *et al.,* 1997; Biondi *et al.,* 1998) have all succeeded in creating active, often fully active, hybrids.

A number of methods have been described for random insertion of short sequences into a target gene, a method commonly referred to as linker scanning mutagenesis. Combinatorial methods for insertion of one domain into another are logical extensions of linker scanning mutagenesis (Fig. 8). Common procedures involve either (1) limited digestion of a target plasmid with DNase I (Heffron *et al.,* 1978) or similar procedures (Luckow *et al.,* 1987) in which the plasmid molecule is randomly linearized at one position followed by ligation of a short linker sequence that often includes a restriction site, or (2) random insertion by transposons (Hallet *et al.,* 1997; Manoil and Bailey, 1997). If the inserted sequence has a restriction enzyme site, a cassette containing the desired gene insert can be cloned into the library of linker insertions. Alternatively, using the DNase I procedure, linkers containing a restriction site can be ligated to the ends of the randomly linearized vector, the linkers digested, and the desired insert cassette ligated to covalently close the vector. This procedure has been employed to randomly insert the green fluorescent protein (GFP) into cAMP-dependent protein kinase regulatory subunit (PKA) in an attempt to develop a viable two-component fluorescence resonance energy transfer (FRET) system for monitoring intracellular cAMP levels (Biondi *et al.,* 1998).

Scanning linker mutagenesis can be used to identify sites amenable to functional insertion of another protein. TEM β-lactamase was inserted into three sites of the *E. coli* maltose binding protein (MalE) (Betton *et al.,* 1997), previously identified by random insertion of a *Bam*H I linker to tolerate small insertions (Duplay *et al.,* 1987). In all three hybrids, both activities were found to be essentially that of the individual wild-type enzymes.

2. *Circular Permutations*

A genetic method for random circular permutation of any gene was first described by Graf and Schachmann (Graf *et al.,* 1996). The method

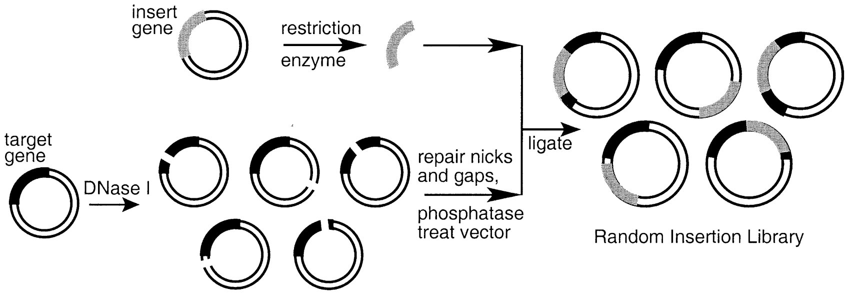


FIG. 8. Random insertion utilizing DNase I treatment (Heffron *et al.,* 1978; Biondi *et al.,* 1998). A plasmid containing the target gene is subjected to DNase I digestion under such conditions that, on average, one double strand break is randomly introduced per molecule. Because DNase I introduces nicks more frequently than double strand breaks and the double strand break does not generally produce blunt ends, repair of the plasmid DNA using a DNA polymerase and DNA ligase is required. The plasmid DNA is then dephosphorylated so that, in the subsequent ligation step, the vector does not religate without insertion of the desired insert fragment. The desired insert fragment is prepared, for example, by suitable restriction digest of the fragment from a second vector. Requirements for this fragment include blunt ends, and removal of the 3′ stop codon. For linker scanning mutagenesis, the insert fragment is a pair of designed oligonucleotides. Due to the random nature of the process, the fragments or oligonucleotide pairs can be inserted in either direction, in any of the three reading frames and in other locations of the plasmid. Thus, the majority of the library members do not have the desired insertions. However, a significant fraction will have an insert within the target gene and be in frame.

is composed of the following steps (Fig. 9): (1) isolation of a linear fragment of double stranded DNA with flanking compatible ends, (2) cyclization of this DNA fragment by ligation under dilute conditions, (3) random linearization of the cyclized gene using DNase I digestion in the presence of Mn^{2+} at dilute concentrations of the enzyme such that the DNase I, on average, makes one double strand break, (4) repair of nicks and gaps using a DNA polymerase and a DNA ligase, and (5) ligation of the fragment into a desired vector by blunt end ligation to create the plasmid library of randomly circularly permuted genes. The vector must be designed so that the 5′ end of the cyclized gene is fused to a start codon and the 3′ is fused to a series of stop codons in all three frames. Using this procedure, one of six members of the library will be both in the correct orientation and the correct frame. Additionally, members of the library can have altered amino acids at either the N- or C-terminal and C-terminal extensions, depending on which stop codon is in frame with the circularly permuted gene. Portions of the experimental protocol described in the original paper (Graf *et*

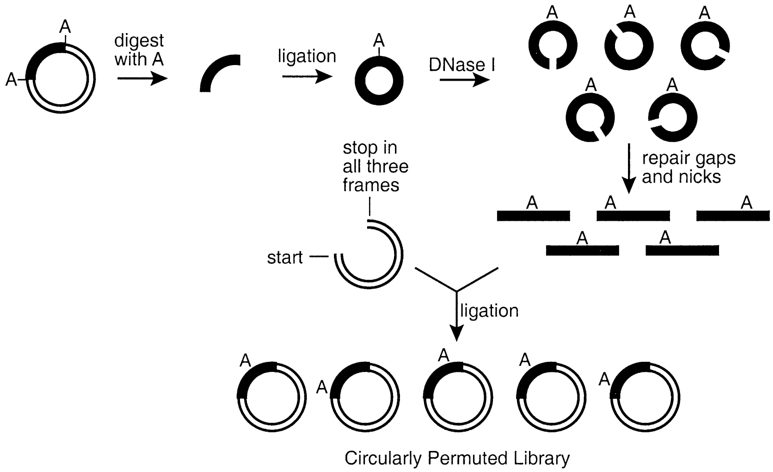


FIG. 9. Random circular permutation (Graf and Schachman, 1996). The gene to be randomly circularly permuted is excised from a suitable vector such that compatible restriction sites exist at the ends. In addition, the ligation of these compatible ends must produce an in-frame fusion of the 5′ and 3′ ends of the gene such that they code for amino acids that will form a suitable "linker" between the original N- and C-terminus of the protein. Ligation under dilute DNA concentrations results in cyclization of the DNA. The cyclized gene is subjected to DNase I digestion under such conditions that, on average, one double strand break is randomly introduced per molecule. As DNase I introduces nicks more frequently than double strand breaks and the double strand break does not generally produce blunt ends, repair of the plasmid DNA using a DNA polymerase and DNA ligase is required. Subsequent ligation into a suitably prepared vector that has dephosphorylated, blunt ends, a start codon, and stop codons in all three frames creates the circularly permuted library in which one out of six members will be both in-frame and in the correct orientation.

al., 1996), particularly the concentration of DNA for circular ligation, do not appear to be correct or are at least suboptimal. The experimental conditions described in a recent paper appear more reasonable (Baird *et al.,* 1999). Thus far, the procedure has been used to systematically identify permissive sites for circular permutation in aspartate transcarbamoylase (Graf *et al.,* 1996), DsbA (Hennecke *et al.,* 1999), and GFP (Baird *et al.,* 1999). Aspartate transcarbamoylase and DsbA tolerated new termini in a variety of locations, whereas GFP was much less permissive.

3. *Incremental Truncation*

Incremental truncation, in its simplest form, allows the creation of a library of every one bp deletion of a gene or gene fragment (Fig. 10). Despite its counterintuitive nature (i.e., that by deleting amino acids

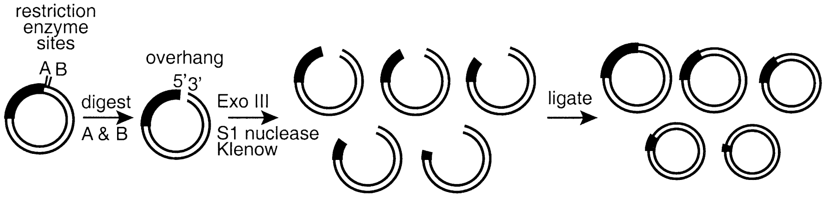


FIG. 10. Incremental truncation libraries (Ostermeier *et al.,* 1999b). Plasmid DNA is digested with two restriction enzymes: one that produces a 3′ recessed end (A; which is susceptible to Exo III digestion) and the other that produces a 5′ recessed end (B; which is resistant to Exo III digestion). Digestion with Exonuclease III proceeds under conditions in which the digestion rate is slow enough so that the removal of aliquots at frequent intervals results in a library of deletions of all possible lengths from one end of the fragment. The ends of the DNA can be blunted by treatment with S1 nuclease and Klenow so that unimolecular ligation results in the desired incremental truncation library.

one can arrive at new function), incremental truncation allows one to explore a number of novel protein engineering strategies (Ostermeier *et al.,* 1999a). A key step in the creation of these libraries is the digestion of the gene fragments with a 3′ to 5′ exonuclease such as Exonuclease III (Exo III) under conditions (e.g., low temperature or in the presence of NaCl) such that the rate is controlled to ~10 bases/min or less. During Exo III digestion, small aliquots are removed frequently and quenched by addition to a low pH, high salt buffer. Blunt ends are prepared by treatment with a single-strand nuclease and a DNA polymerase followed by unimolecular ligation to recyclize the vector. As Exo III digests DNA at a substantially uniform and synchronous rate (Wu *et al.,* 1976), this allows the creation of a library of every one bp deletion of a gene or gene fragment.

Incremental truncation libraries can be used to examine all possible bisection points within a given region of an enzyme that will allow the conversion of a monomeric enzyme into its functional heterodimer (Ostermeier *et al.,* 1999b). This strategy is described in the lower left panel of Figure 11 when the two starting fragments are overlapping fragments of the same gene. Conversion of a monomeric enzyme to a heterodimer by breaking a link in the peptide backbone can be considered ‘‘reverse evolution’’ since such a process is the reverse of domain recruitment. As such, it may identify ancestral fusion points and be an experimental approach to functionally defining domain boundaries.

The *E. coli* GARTase PurN has been systematically bisected by this method and found to tolerate bisection in two regions: the first in a nonconserved region in the vicinity of a domain boundary suggested by sequence alignments, and the second in a highly conserved region three residues away from an active site residue (Ostermeier *et al.,* 1999b). Bisection in the second region indicated that it was possible to divide the active site residues onto two separate polypeptides and that two residues conserved across all GARTases and other formyltransferases are non-essential in the context of a heterodimer. Subsequent work on the creation of fusion libraries between *E. coli* and human GARTases confirmed the second region, but not the first, as a suitable junction for domain swapping, at least within GARTases (Ostermeier *et al.,* 1999c). However, the suitability of regions in the human enzyme for functional bisection has not been examined. It would be interesting to see if the intersection of the sets of functional bisection sites in the two enzymes would predict locations for successful domain swapping between them.

Incremental truncation can also be used to create a library of fusions between all possible fragments of two genes using a method called incremental truncation for the creation of hybrid enzymes (ITCHY) (Fig. 11). Overlapping fragments of two genes are cloned into suitable vectors for incremental truncation. As shown in the lower right panel of Figure 11, incremental truncation is performed as before, but the judicious location of a restriction site after the 3′ end of the C-terminal gene allows the fusion of one incremental truncation library to the other by ligation. Since this ligation is a random process, all possible fragments of one gene will be fused to all possible fragments of the other gene, provided a sufficient library size is obtained (which depends on the product of the maximum number of bases truncated in each fragment). The ITCHY library contains fusions between genes where they align, as well as internal ‘‘duplications’’ and ‘‘deletions.’’ The size range of ‘‘duplications’’ and ‘‘deletions’’ can be selected, for example, by digesting the library with 5′ and 3′ restriction sites, by separation by gel electrophoresis, and by subcloning the desired size. In this manner, the library can be selected to have fusions near the locations where the sequences or structures align. ITCHY performed on a single gene will create a library of internal tandem duplications and deletions of that gene.

ITCHY has been used to create functional interspecies hybrids between the *E. coli* and human genes for GARTase, genes that have low DNA sequence homology (50% identity on the DNA level) (Ostermeier *et al.,* 1999c). ITCHY was found to identify a more diverse set of active hybrids than DNA shuffling. Furthermore, the most active hybrid, fused at a region of non-DNA homology, was solely identified by ITCHY. All fusion points of active enzymes except one were exactly at the alignment of the two genes. This was somewhat surprising, as it was reasonable to

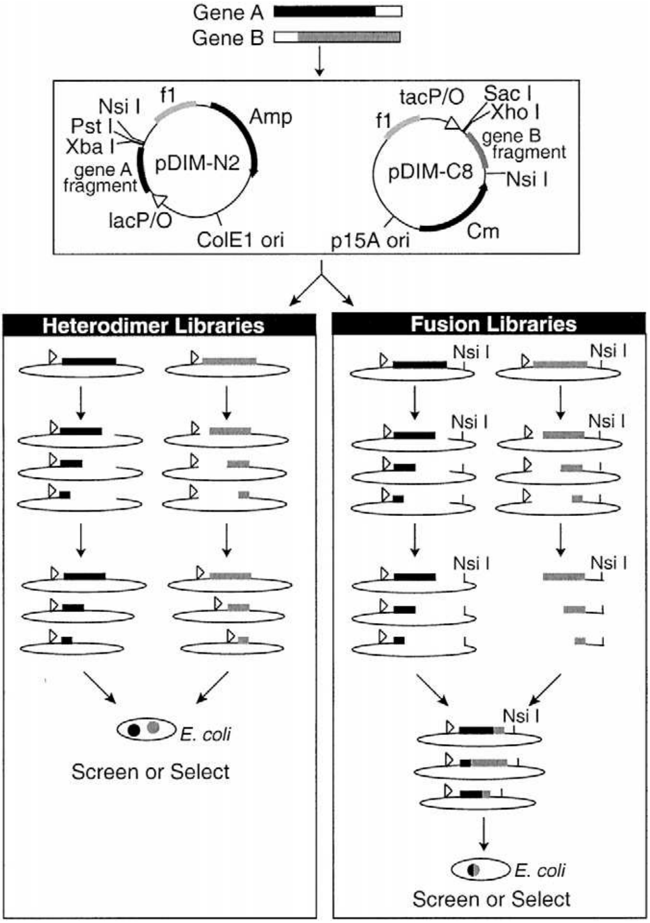


FIG. 11. Incremental truncation for the creation of hybrid enzymes (ITCHY) (Ostermeier *et al.,* 1999a; Ostermeier *et al.,* 1999c). A large 5′ fragment of gene A is cloned into phagemid pDIM-N2 and a large 3′ fragment of gene B is cloned into phagemid pDIM-C8. Phagemids pDIM-N2 and pDIM-C8 contain different antibiotic resistance genes (Amp, Cm) and different origins of replication (ColE1 and p15A) so that both can be maintained in the same cell, if desired. The phagemids also have restriction sites designed for creating incremental truncation libraries from the 3′ end of the gene fragment in pDIM-N2 and the 5′ end of the gene fragment in pDIM-C8. Noncovalent ITCHY is described in the lower left panel. Incremental truncation is performed on each plasmid and the vectors circularized as in Figure 9 such that a series of stop codons in all three frames is fused to the 3′ end of truncations of gene A and a start codon is fused to the 5′ end of truncations of gene B. Both vectors can be transformed into the same cell to create a library of (potentially) heterodimeric enzymes. If gene A and gene B are the same gene, this method allows one to search for loci that allow functional bisection

expect that fusions would tolerate a few extra amino acids or a few deleted amino acids, particularly in loops. However, the linear distances between conserved residues may have some importance for structure and/or function. Alternatively, the decrease in activity caused by extra or deleted amino acids may be small, but significant enough to inactivate a fusion protein that has an activity (as judged by k_{cat}/ K_m) 500 to 10,000 less than wild type. Regardless, this work shows that ITCHY is a combinatorial solution to generating active fusion proteins that would be difficult to predict *a priori,* as most active hybrids found were fused in the proximity of the active site.

4. *Internal Tandem Duplications or Deletions*

A strategy utilizing *Bal1*31 nuclease to create deletions, originally developed to investigate the arabinose promoter region (*araBAD*) (Reeder and Schleif, 1993), has been used to examine the tolerance of the linker region of the AraC protein for insertions and deletions (Eustance *et al.,* 1994) and of staphylococcal nuclease for internal tandem duplications (Nguyen and Schleif, 1998). *Bal*31 nuclease is a double-stranded nuclease that has both 3′ to 5′ and 5′ to 3′ exonuclease activities. A plasmid constructed of overlapping fragments of the same gene separated by a unique restriction site was linearized at this site, digested for various lengths of time with *Bal*31 nuclease, and recircularized. Although such libraries have not been characterized in great detail, they should be heavily biased toward having fusions between fragments that have had the same amount of bp deleted from them. Deviation from the bias would only result from variations within the sample between the amounts of DNA digested from each end (Eustance *et al.,* 1994). However, the inability to protect one end of DNA from *Bal*31 nuclease digestion precludes using it for creating incremental truncation-like libraries that are fused to defined sequences, such as start codons and stop codons, or for use in creating ITCHY libraries, without the use of a very long spacer.

VI. Perspective

Progress in utilizing domain swapping for protein engineering has been partially hindered by the lack of combinatorial methods for per-

←

of a protein (Ostermeier *et al.,* 1999b) Covalent ITCHY is described in the lower right panel. Incremental truncation is followed by digestion with restriction enzyme *Nsi*I, isolation of the indicated DNA fragments, and ligation to form a library of gene fusions. The resulting fusion library has random fragments of the N terminus of protein A fused to random fragments of the C terminus of protein B. If gene A and gene B are the same gene, this method allows one to create a library of internal deletions and tandem duplications of a gene.

forming domain swapping on any two proteins independent of DNA homology. Often decisions made on where to fuse the enzymes are made irrespective of domain boundary considerations in lieu of the location of natural restrictions sites in the two genes that allow the construction of a set of hybrid genes. Recently developed combinatorial methods such as random insertion of large domains (Biondi *et al.,* 1998), random circular permutation (Graf and Schachman, 1996), and incremental truncation (Ostermeier *et al.,* 1999a; Ostermeier *et al.,* 1999c) should stimulate progress in the field.

DNA shuffling can combinatorialize related genes based on DNA homology. However, DNA shuffling's success in the evolution of novel function derives from the combinatorializing of homologous structures. Molecular breeding (i.e., DNA shuffling of families of proteins) can evolve new function more efficiently by accessing a more diverse sequence space, but in a rational manner: staying within homologous structures. This bodes well for a strategy of protein engineering that incorporates domain swapping of homologous structures. The high activity and crystal structure of domain-swapped fusions between trypsin and fXa, which have 38% sequence identity on the amino-acid level, illustrates how tolerant homologous structures can be in accommodating mismatches of proteins with low sequence homology (Hopfner *et al.,* 1998). Indeed, it is likely that such plasticity has been evolutionarily programmed into proteins to enable the natural evolution of new function. Continuing to explore the natural evolution of proteins is likely to be important for advances in protein engineering.

Acknowledgments

We thank Dr. Stefan Lutz for many useful comments in the preparation of this manuscript.

References

Aimi, J., Qiu, H., Williams, J., Zalkin, H., and Dixon, J. E. (1990). *De novo* purine nucleotide biosynthesis: cloning of human and avian cDNAs encoding the trifunctional glycinamide ribonucleotide synthetase-aminoimidazole ribonucleotide synthetase-glycinamide ribonucleotide transformylase by functional complementation in *E. coli. Nucleic Acids Res.,* **18,** 6665–6672.

Almassy, R. J., Janson, C. A., Kan, C.-C., and Hostomska, Z. (1992). Structures of apo and complexed *Escherichia coli* glycinamide ribonucleotide transformylase. *Proc. Natl. Acad. Sci. USA,* **89,** 6114–6118.

Arnold, F. H., and Volkov, A. A. (1999). Directed evolution of biocatalysts. *Curr. Opin. Chem. Biol,* **3,** 54–59.

Baird, G. S., Zacharias, D. A., and Tsien, R. Y. (1999). Circular permutation and receptor insertion within green fluorescent proteins. *Proc. Natl. Acad. Sci. USA,* **96,** 11241–11246.

Baneyx, F. (1999). Recombinant protein expression in *Escherichia coli. Curr. Opin. Biotechnol.,* **10,** 411–421.

Baron, M., Norman, D. G., and Campbell, I. D. (1991). Protein modules. *Trends Biochem. Sci.,* **16,** 13–17.

Beguin, P. (1999). Hybrid enzymes. *Curr. Opin. Biotechnol.,* **10,** 336–340.

Belkaïd, M., Penverne, B., Denis, M., and Hervé, G. (1987). *In situ* behavior of the pyrimidine pathway enzymes in *Saccharomyces cerevisiae.* 2. Reaction mechanism of aspartate transcarbamylase dissociated from carbamylphosphate synthetase by genetic alteration. *Arch. Biochem. Biophys.,* **254,** 568–578.

Betton, J.-M., Jacob, J. P., Hofnung, M., and Broome-Smith, J. K. (1997). Creating a bifunctional protein by insertion of β-lactamase into the maltodextrin-binding protein. *Nat. Biotechnology,* **15,** 1276–1279.

Biondi, R. M., Baehler, P. J., Reymond, C. D., and Véron, M. (1998). Random insertion of GFP into the cAMP-dependent protein kinase regulatory subunit from *Dictyostelium discoideum. Nucleic Acids Res.,* **26,** 4946–4952.

Bulow, L. (1998). Preparation of artificial bifunctional enzymes by gene fusion. *Biochem. Soc. Symp.,* **57,** 123–133.

Caffrey, P., Bevitt, D. J., Staunton, J., and Leadlay, P. F. (1992). Identification of DEBS 1, DEBS 2 and DEBS 3, the multienzyme polypeptides of the erythromycin-producing polyketide synthase from *Saccharopolyspora erythraea. FEBS Lett.,* **304,** 225–228.

Carreras, C. W., and Santi, D. V. (1998). Engineering of modular polyketide synthases to produce novel polyketides. *Curr. Opin. Biotechnol.,* **9,** 403–411.

Chandrasegaran, S., and Smith, J. (1999). Chimeric restriction enzymes: what is next? *Biol. Chem.,* **380,** 841–848.

Clackson, T. (1997). Controlling mammalian gene expression with small molecules. *Curr. Opin. Chem. Biol,* **1,** 210–218.

Cortes, J., Haydock, S. F., Roberts, G. A., Bevitt, D. J., and Leadlay, P. F. (1990). An unusually large multifunctional polypeptide in the erythromycin-producing polyketide synthase of *Saccharopolyspora erythraea. Nature,* **348,** 176–178.

Cozier, G. E., and Anthony, C. (1995a). Structure of the quinoprotein glucose dehydrogenase of *Escherichia coli* modelled on that of methanol dehydrogenase from *Methylobacterium extorquens. Biochem. J.,* **312,** 679–685.

Cozier, G. E., Giles, I. G., and Anthony, C. (1995b). The structure of the quinoprotein alcohol dehydrogenase of Acetobacter aceti modelled on that of methanol dehydrogenase from Methylobacterium extorquens. *Biochem. J.,* **308,** 375–379.

Crameri, A., Raillard, S.-A., Bermudez, E., and Stemmer, W. P. C. (1998). DNA shuffling of a family of genes from diverse species accelerates directed evolution. *Nature,* **391,** 288–291.

Davidson, J. N., Chen, K. C., Jamison, R. S., Musmanno, L. A., and Kern, C. B. (1993). The evolutionary history of the first three enzymes in pyrimidine biosynthesis. *BioEssays,* **15,** 157–164.

Davidson, J. N., and Peterson, M. L. (1997). Origin of genes encoding multi-enzymatic proteins in eukaryotes. *Trends Genet.,* **13,** 281–285.

de Ferra, F., Rodriguez, F., Tortora, O., Tosi, C., and Grandi, G. (1997). Engineering of peptide synthetases: Key role of the thioesterase-like domain for efficient production of recombinant peptides. *J. Biol. Chem.,* **272,** 25304–25309.

Doi, N., Itaya, M., Yomo, T., Tokura, S., and Yanagawa, H. (1997). Insertion of foreign random sequences of 120 amino acid residues into an active enzyme. *FEBS Lett.,* **402,** 177–180.

Doi, N., and Yanagawa, H. (1999). Design of generic biosensors based on green fluorescent proteins with allosteric sites by directed evolution. *FEBS Lett.,* **453,** 305–307.

Doolittle, R. F., and Bork, P. (1993). Evolutionarily mobile modules in proteins. *Sci. Am.,* **269,** 50–56.

Drees, B. L. (1999). Progress and variations in two-hybrid and three-hybrid technologies. *Curr. Opin. Chem. Biol,* **3,** 64–70.

Duplay, P., Szmelcman, S., Bedouelle, H., and Hofnung, M. (1987). Silent and functional changes in the periplasmic maltose-binding protein of *Escherichia coli* K12. I. Transport of maltose. *J. Mol. Biol.,* **194,** 663–673.

Durbecq, V., Legrain, C., Roovers, M., Pierard, A., and Glansdorff, N. (1997). The carbamate kinase-like carbamoyl phosphate synthetase of the hyperthermophilic archaeon *Pyrococcus furiosus,* a missing link in the evolution of carbamoyl phosphate biosynthesis. *Proc. Natl. Acad. Sci. USA,* **94,** 12803–12808.

Ebbole, D. J., and Zalkin, H. (1987). Cloning and characterization of a 12-gene cluster from *Bacillus subtilis* encoding nine enzymes for de novo purine nucleotide synthesis. *J. Biol. Chem.,* **262,** 8274–8287.

Eustance, R. J., Bustos, S. A., and Schleif, R. F. (1994). Reaching out. Locating and lengthening the interdomain linker in AraC protein. *J. Mol. Biol.,* **242,** 330–338.

Feng, Y., Minnerly, J. C., Zurfluh, L. L., Joy, W. D., Hood, W. F., Abegg, A. L., Grabbe, E. S., Shieh, J. J., Thurman, T. L., McKearn, J. P., and McWherter, C. A. (1999). Circular permutation of granulocyte colony-stimulating factor. *Biochemistry,* **38,** 4553–4563.

Fushinobu, S., Ohta, T., and Matsuzawa, H. (1998). Homotropic activation via the subunit interaction and allosteric symmetry revealed on analysis of hybrid enzymes of L-lactate dehydrogenase. *J. Biol. Chem.,* **273,** 2971–2976.

Georgiou, G., Stathopoulus, C., Daugherty, P. S., Nayak, A. R., Iverson, B. L., and Curtiss III, R. (1997). Display of heterologous proteins on the surface of microorganisms: from the screening of combinatorial libraries to live recombinant vaccines. *Nature Biotechnology,* **15,** 29–34.

Ghosh, M., Anthony, C., Harlos, K., Goodwin, M. G., and Blake, C. (1995). The refined structure of the quinoprotein methanol dehydrogenase from *Methylobacterium extorquens* at 1.94 A. *Structure,* **3,** 177–187.

Goldenberg, D. P., and Creighton, T. E. (1983). Circular and circularly permuted forms of bovine pancreatic trypsin inhibitor. *J. Mol. Biol.,* **165,** 407–413.

Graf, R., and Schachman, H. K. (1996). Random circular permutation of genes and expressed polypeptide chains: Application of the method to the catalytic chains of aspartate transcarbamoylase. *Proc. Natl. Acad. Sci. USA,* **93,** 11591–11596.

Grandori, R., Struck, K., Giovanielli, K., and Carey, J. (1997). A three-step PCR protocol for construction of chimeric proteins. *Protein Eng.,* **10,** 1099–1100.

Greisman, H. A., and Pabo, C. O. (1997). A general strategy for selecting high-affinity zinc finger proteins for diverse DNA target sites. *Science,* **275,** 657–661.

Guddat, L. W., Bardwell, J. C., and Martin, J. L. (1998). Crystal structures of reduced and oxidized DsbA: investigation of domain motion and thiolate stabilization. *Structure,* **6,** 757–767.

Guruprasad, K., Tormakangas, K., Kervinen, J., and Blundell, T. L. (1994). Comparative modelling of barley-grain aspartic proteinase: a structural rationale for observed hydrolytic specificity. *FEBS Lett.,* **352,** 131–136.

Guy, H. I., and Evans, D. R. (1996). Function of the major synthetase subdomains of carbamyl-phosphate synthetase. *J. Biol. Chem.,* **271,** 13762–13769.

Guy, H. I., Schmidt, B., Hervé, G., and Evans, D. R. (1998). Pressure-induced dissociation of carbamoyl-phosphate synthetase domains. The catalytically active form is dimeric. *J. Biol. Chem.,* **273,** 14172–14178.

Hallet, B., Sherratt, D. J., and Hayes, F. (1997). Pentapeptide scanning mutagenesis: random insertion of a variable five amino acid cassette in a target protein. *Nucleic Acids Res.,* **25,** 1866–1867.

Hawkins, A. R., and Lamb, H. K. (1995). The molecular biology of multidomain proteins: selected examples. *Eur. J. Biochem.,* **232,** 7–18.

Heffron, F., So, M., and McCarthy, B. J. (1978). In vitro mutagenesis of a circular DNA molecule by using synthetic restriction sites. *Proc. Natl. Acad. Sci. USA,* **75,** 6012–6016.

Hegyi, H., and Bork, P. (1997). On the classification and evolution of protein modules. *J. Protein Chem.,* **16,** 545–551.

Heinemann, U., and Hahn, M. (1995). Circular permutation of polypeptide chains: implications for protein folding and stability. *Prog. Biophys. Mol. Biol,* **64,** 121–143.

Henikoff, S. (1986). The *Saccharomyces cerevisiae* ADE5,7 protein is homologous to overlapping *Drosophila melanogaster* Gart polypeptides. *J. Mol. Biol.,* **190,** 519–528.

Henikoff, S., Keene, M. A., Sloan, J. S., Bleskan, J., Hards, R., and Patterson, D. (1986). Multiple purine pathway enzyme activities are encoded at a single genetic locus in *Drosophila. Proc. Natl. Acad. Sci. USA,* **83,** 720–724.

Hennecke, J., Sebbel, P., and Glockshuber, R. (1999). Random circular permutation of DsbA reveals segments that are essential for protein folding and stability. *J. Mol. Biol.,* **286,** 1197–1215.

Hervé, G., Nagy, M., Le Gouar, M., Penverne, B., and Ladjimi, M. (1993). The carbamoyl phosphate synthetase–aspartate transcarbamoylase complex of *Saccharomyces cerevisiae:* molecular and cellular aspects. *Biochem. Soc. Trans.,* **21,** 195–198.

Hoogenboom, H. R., de Bruine, A. P., Hufton, S. E., Hoet, R. M., Arends, J. W., and Roovers, R. C. (1998). Antibody phage display technology and its applications. *Immunotechnology,* **4,** 1–20.

Hopfner, K.-P., Kopetzki, E., Kreb(German B)e, G.-B., Bode, W., Huber, R., and Engh, R. A. (1998). New enzyme lineages by subdomain shuffling. *Proc. Natl. Acad. Sci. USA,* **95,** 9813–9818.

Horton, R. M., Hunt, H. D., Ho, S. N., Pullen, J. K., and Pease, L. R. (1989). Engineered hybrid genes without the use of restriction enzymes: gene splicing by overlap extension. *Gene,* **77,** 61–68.

Huang, B., Schaeffer, C. J., Li, Q., and Tsai, M. D. (1996). Splase: a new class IIS zinc-finger restriction endonuclease with specificity for Sp1 binding sites. *J. Protein Chem.,* **15,** 481–489.

Hutchinson, C. R. (1999). Microbial polyketide synthases: more and more prolific. *Proc. Natl. Acad. Sci. USA,* **96,** 3336–3338.

Isalan, M., Choo, Y., and Klug, A. (1997). Synergy between adjacent zinc fingers in sequence-specific DNA recognition. *Proc. Natl. Acad. Sci. USA,* **94,** 5617–5621.

Jeltsch, A. (1999). Circular permutations in the molecular evolution of DNA methyltransferases. *J. Mol. Evol.,* **49,** 161–164.

Kan, J. L., Jannatipour, M., Taylor, S. M., and Moran, R. G. (1993). Mouse cDNAs encoding a trifunctional protein of *de novo* purine synthesis and a related single-domain glycinamide ribonucleotide synthetase. *Gene,* **137,** 195–202.

Kao, C. M., Mcpherson, M., McDaniel, R. N., Fu, H., Cane, D. E., and Khosla, C. (1997). Gain of function mutagenesis of the erythromycin polyketide synthase. 2. Engineered biosynthesis of an eight-membered ring tetraketide lactone. *J. Am. Chem. Soc.,* **119,** 11339–11340.

Kao, C. M., Mcpherson, M., McDaniel, R. N., Fu, H., Cane, D. E., and Khosla, C. (1998). Alcohol stereochemistry in polyketide backbones is controlled by the β-ketoreductase domains of modular polyketide synthases. *J. Am. Chem. Soc.,* **120,** 2478–2479.

Kapust, R. B., and Waugh, D. S. (1999). *Escherichia coli* maltose-binding protein is uncommonly effective at promoting the solubility of polypeptides to which it is fused. *Protein Sci.,* **8,** 1668–1674.

Keating, T. A., and Walsh, C. T. (1999). Initiation, elongation, and termination strategies in polyketide and polypeptide antibiotic biosynthesis. *Curr. Opin. Chem. Biol.,* **3,** 598–606.

Kikuchi, M., Ohnishi, K., and Harayama, S. (1999). Novel family shuffling methods for the *in vitro* evolution of enzymes. *Gene,* **236,** 159–167.

Kim, S. C., Posfai, G., and Szybalski, W. (1991). A novel gene-fusing vector: construction of a 5′-GGmCC-specific chimeric methyltransferase, M-*Bsp*RI/M-*Bsu*RI. *Gene,* **100,** 45–50.

Kim, Y. G., and Chandrasegaran, S. (1994). Chimeric restriction endonuclease. *Proc. Natl. Acad. Sci. USA,* **91,** 883–887.

Kim, Y. G., Cha, J., and Chandrasegaran, S. (1996). Hybrid restriction enzymes: zinc finger fusions to *Fok*I cleavage domain. *Proc. Natl. Acad. Sci. USA,* **93,** 1156–1160.

Kim, Y. G., Shi, Y., Berg, J. M., and Chandrasegaran, S. (1997). Site-specific cleavage of DNA-RNA hybrids by zinc finger/*Fok*I cleavage domain fusions. *Gene,* **203,** 43–49.

Kim, Y. G., Smith, J., Durgesha, M., and Chandrasegaran, S. (1998). Chimeric restriction enzyme: Gal4 fusion to *Fok*I cleavage domain. *Biol. Chem.,* **379,** 489–495.

Kreitman, R. J., Puri, R. K., and Pastan, I. (1994). A circularly permuted recombinant interleukin 4 toxin with increased activity. *Proc. Natl. Acad. Sci. USA,* **91,** 6889–6893.

Kreitman, R. J., Puri, R. K., and Pastan, I. (1995). Increased antitumor activity of a circularly permuted interleukin 4- toxin in mice with interleukin 4 receptor-bearing human carcinoma. *Cancer Res.,* **55,** 3357–3363.

Kuhstoss, S., Huber, M., Turner, J. R., Paschal, J. W., and Rao, R. N. (1996). Production of a novel polyketide through the construction of a hybrid polyketide synthase. *Gene,* **183,** 231–236.

Labedan, B., Boyen, A., Baetens, M., Charlier, D., Chen, P., Cunin, R., Durbeco, V., Glansdorff, N., Herve, G., Legrain, C., Liang, Z., Purcarea, C., Roovers, M., Sanchez, R., Toong, T. L., Van de, C. M., van Vliet, F., Xu, Y., and Zhang, Y. F. (1999). The evolutionary history of carbamoyltransferases: A complex set of paralogous genes was already present in the last universal common ancestor. *J. Mol. Evol.,* **49,** 461–473.

Ladant, D., Glaser, P., and Ullmann, A. (1992). Insertional mutagenesis of *Bordetella pertussis* adenylate cyclase. *J. Biol. Chem.,* **267,** 2244–2250.

Lawrence, J. G. (1997). Selfish operons and speciation by gene transfer. *Trends Microbiol.,* **5,** 355–359.

Li, L., Wu, L. P., and Chandrasegaran, S. (1992). Functional domains in *Fok*I restriction endonuclease. *Proc. Natl. Acad. Sci. USA,* **89,** 4275–4279.

Li, L., Wu, L. P., Clarke, R., and Chandrasegaran, S. (1993). C-terminal deletion mutants of the *Fok*I restriction endonuclease. *Gene,* **133,** 79–84.

Lim, F., Morris, C. P., Occhiodoro, F., and Wallace, J. C. (1988). Sequence and domain structure of yeast pyruvate carboxylase. *J. Biol. Chem.,* **263,** 11493–11497.

Lindbladh, C., Persson, M., Bulow, L., and Mosbach, K. (1992). Characterization of a recombinant bifunctional enzyme, galactose dehydrogenase/bacterial luciferase, displaying an improved bioluminescence in a three-enzyme system. *Eur. J. Biochem.,* **204,** 241–247.

Lindbladh, C., Rault, M., Hagglund, C., Small, W. C., Mosbach, K., Bulow, L., Evans, C., and Srere, P. A. (1994). Preparation and kinetic characterization of a fusion protein of yeast mitochondrial citrate synthase and malate dehydrogenase. *Biochemistry,* **33,** 11692–11698.

Lindqvist, Y., and Schneider, G. (1997). Circular permutations of natural protein sequences: structural evidence. *Curr. Opin. Struct. Biol,* **7,** 422–427.

Liu, L., Thanchaipenet, A., Fu, H., Betlach, M., and Ashley, G. (1997). Biosynthesis of 2-nor-6-deoxyerythronolide B by rationally designed domain substitutions. *J. Am. Chem. Soc.,* **119,** 10553–10554.

Ljungcrantz, P., Carlsson, H., Mansson, M. O., Buckel, P., Mosbach, K., and Bulow, L. (1989). Construction of an artificial bifunctional enzyme, beta-galactosidase/galactose dehydrogenase, exhibiting efficient galactose channeling. *Biochemistry,* **28,** 8786–8792.

Luckow, B., Renkawitz, R., and Schutz, G. (1987). A new method for constructing linker scanning mutants. *Nucleic Acids Res.,* **15,** 417–429.

Mally, M. I., Grayson, D. R., and Evans, D. R. (1980). Catalytic synergy in the multifunctional protein that initiates pyrimidine biosynthesis in Syrian hamster cells. *J. Biol Chem.,* **255,** 11372–11380.

Manoil, C., and Bailey, J. (1997). A simple screen for permissive sites in proteins: analysis of *Escherichia coli* lac permease. *J. Mol. Biol.,* **267,** 250–263.

Marsden, A. F., Wilkinson, B., Cortes, J., Dunster, N. J., Staunton, J., and Leadlay, P. F. (1998). Engineering broader specificity into an antibiotic-producing polyketide synthase. *Science,* **279,** 199–202.

Martin, J. L., Bardwell, J. C., and Kuriyan, J. (1993). Crystal structure of the DsbA protein required for disulphide bond formation *in vivo. Nature,* **365,** 464–468.

Martoglio, B., and Dobberstein, B. (1998). Signal sequences: more than just greasy peptides. *Trends Cell. Biol.,* **8,** 410–415.

McDaniel, R., Kao, C. M., Fu, H., Hevezi, P., Gustafsson, C., Betlach, M., Ashley, G., Cane, D. E., and Khosla, C. (1997). Gain-of-function mutagenesis of a modular polyketide synthase. *J. Am. Chem. Soc.,* **119,** 4309–4310.

McDaniel, R., Thamchaipenet, A., Gustafsson, C., Fu, H., Betlach, M., and Ashley, G. (1999). Multiple genetic modifications of the erythromycin polyketide synthase to produce a library of novel ''unnatural'' natural products. *Proc. Natl. Acad. Sci. USA,* **96,** 1846–1851.

McLachlan, A. D. (1979). Gene duplications in the structural evolution of chymotrypsin. *J. Mol. Biol.,* **128,** 49–79.

McLachlan, A. D. (1987). Gene duplication and the origin of repetitive protein structures. *Cold Spring Harbor Symp. Quant. Biol.,* **52,** 411–420.

McWherter, C. A., Feng, Y., Zurfluh, L. L., Klein, B. K., Baganoff, M. P., Polazzi, J. O., Hood, W. F., Paik, K., Abegg, A. L., Grabbe, E. S., Shieh, J. J., Donnelly, A. M., and McKearn, J. P. (1999). Circular permutation of the granulocyte colony-stimulating factor receptor agonist domain of myelopoietin. *Biochemistry,* **38,** 4564–4571.

Miyawaki, A., Llopis, J., Heim, R., McCaffery, J. M., Adams, J. A., Ikura, M., and Tsien, R. Y. (1997). Fluorescent indicators for Ca2+ based on green fluorescent proteins and calmodulin. *Nature,* **388,** 882–887.

Mootz, H. D., and Marahiel, M. A. (1999). Design and application of multimodular peptide synthetases. *Curr. Opin. Biotechnol.,* **10,** 341–348.

Netzer, W. J., and Hartl, F. U. (1997). Recombination of protein domains facilitated by co-translational folding in eukaryotes. *Nature,* **388,** 343–349.

Nguyen, D. M., and Schleif, R. F. (1998). Isolation and physical characterization of random insertions in *Staphylococcal nuclease. J. Mol. Biol.,* **282,** 751–759.

Nilsson, J., Stahl, S., Lundeberg, J., Uhlen, M., and Nygren, P.-A. (1997). Affinity fusion strategies for detection, purification, and immobilization of recombinant proteins. *Protein Expression Purif.,* **11,** 1–16.

Nixon, A. E., Warren, M. S., and Benkovic, S. J. (1997). Assembly of an active enzyme by the linkage of two protein modules. *Proc. Natl. Acad. Sci. USA,* **94,** 1069–1073.

Nixon, A. E., Ostermeier, M., and Benkovic, S. J. (1998). Hybrid enzymes: manipulating enzyme design. *Trends Biotechnol.,* **16,** 258–264.

Nyunoya, H., and Lusty, C. J. (1983). The carB gene of *Escherichia coli:* a duplicated gene coding for the large subunit of carbamoyl-phosphate synthetase. *Proc. Natl. Acad. Sci. USA,* **80,** 4629–4633.

Oliynyk, M., Brown, M. J. B., Cortes, J., Staunton, J., and Leadlay, P. F. (1996). A hybrid modular polyketide synthase obtained by domain swapping. *Chem. Biol.,* **3,** 833–839.

Olsen, O., Thomsen, K. K., Weber, J., Duus, J. O., Svendsen, I., Wegener, C., and von Wettstein, D. (1996). Transplanting two unique beta-glucanase catalytic activities into one multienzyme, which forms glucose. *Biotechnology* (*NY*), **14,** 71–76.

Ostermeier, M., Nixon, A. E., and Benkovic, S. J. (1999a). Incremental truncation as a strategy in the engineering of novel biocatalysts. *Bioorg. Med. Chem.,* **7,** 2139–2144.

Ostermeier, M., Nixon, A. E., Shim, J. H., and Benkovic, S. J. (1999b). Combinatorial protein engineering by incremental truncation. *Proc. Natl. Acad. Sci. USA,* **96,** 3562–3567.

Ostermeier, M., Shim, J. H., and Benkovic, S. J. (1999c). A combinatorial approach to hybrid enzymes independent of DNA homology. *Nat. Biotechnol.,* **17,** 1205–1209.

Pavletich, N. P., and Pabo, C. O. (1991). Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 A. *Science,* **252,** 809–817.

Pelletier, J. N., Campbell-Valois, F. X., and Michnick, S. W. (1998). Oligomerization domain-directed reassembly of active dihydrofolate reductase from rationally designed fragments. *Proc. Natl. Acad. Sci. USA,* **95,** 12141–12146.

Perona, J. J., and Craik, C. S. (1997). Evolutionary divergence of substrate specificity within the chymotrypsin-like serine protease fold. *J. Biol. Chem.,* **272,** 29987–29990.

Ponting, C. P., and Russell, R. B. (1995). Swaposins: circular permutations within genes encoding saposin homologues. *Trends Biochem. Sci.,* **20,** 179–180.

Purcarea, C., Simon, V., Prieur, D., and Hervé, G. (1996). Purification and characterization of carbamoyl-phosphate synthetase from the deep-sea hyperthermophilic archaebacterium *Pyrococcus abyssi. Eur. J. Biochem.,* **236,** 189–199.

Purcarea, C., Evans, D. R., and Hervé, G. (1999). Channeling of carbamoyl phosphate to the pyrimidine and arginine biosynthetic pathways in the deep sea hyperthermophilic archaeon *Pyrococcus abyssi. J. Biol. Chem.,* **274,** 6122–6129.

Quemeneur, E., Moutiez, M., Charbonnier, J. B., and Menez, A. (1998). Engineering cyclophilin into a proline-specific endopeptidase. *Nature,* **391,** 301–304.

Reeder, T., and Schleif, R. (1993). AraC protein can activate transcription from only one position and when pointed in only one direction. *J. Mol. Biol.,* **231,** 205–218.

Reiter, Y., and Pastan, I. (1998). Recombinant Fv immunotoxins and Fv fragments as novel agents for cancer therapy and diagnosis. *Trends Biotechnol.,* **16,** 513–520.

Remy, I., and Michnick, S. W. (1999). Clonal selection and *in vivo* quantitation of protein interactions with protein-fragment complementation assays. *Proc. Natl. Acad. Sci. USA,* **96,** 5394–5399.

Ritsema, T., Gehring, A. M., Stuitje, A. R., van der Drift, K. M., Dandal, I., Lambalot, R. H., Walsh, C. T., Thomas-Oates, J. E., Lugtenberg, B. J., and Spaink, H. P. (1998). Functional analysis of an interspecies chimera of acyl carrier proteins indicates a specialized domain for protein recognition. *Mol. Gen. Genet.,* **257,** 641–648.

Rodi, D. J., and Makowski, L. (1999). Phage-display technology—finding a needle in a vast molecular haystack. *Curr. Opin. Biotechnol.,* **10,** 87–93.

Romoser, V. A., Hinkle, P. M., and Persechini, A. (1997). Detection in living cells of Ca2+-dependent changes in the fluorescence emission of an indicator composed of two green fluorescent protein variants linked by a calmodulin-binding sequence. A new class of fluorescent indicators. *J. Biol. Chem.,* **272,** 13270–13274.

Rossmann, M. G., Moras, D., and Olsen, K. W. (1974). Chemical and biological evolution of nucleotide-binding protein. *Nature,* **250,** 194–199.

Ruan, X., Pereda, A., Stassi, D. L., Zeidner, D., Summers, R. G., Jackson, M., Shivakumar, A., Kakavas, S., Staver, M. J., Donadio, S., and Katz, L. (1997). Acyltransferase domain substitutions in erythromycin polyketide synthase yield novel erythromycin derivatives. *J. Bacteriol.,* **179,** 6416–6425.

Russell, R. B. (1994). Domain insertion. *Protein Eng.,* **7,** 1407–1410.

Russell, R. B., and Ponting, C. P. (1998). Protein fold irregularities that hinder sequence analysis. *Curr. Opin. Struct. Biol.,* **8,** 364–371.

Schlunegger, M. P., Bennett, M. J., and Eisenberg, D. (1997). Oligomer formation by 3D domain swapping: a model for protein assembly and misassembly. *Adv. Protein Chem.,* **50,** 61–122.

Schmitt, E., Blanquet, S., and Mechulam, Y. (1996). Structure of crystalline *Escherichia coli* methionyl-tRNA(f)Met formyltransferase: comparison with glycinamide ribonucleotide formyltransferase. *EMBO J.,* **15,** 4749–4758.

Schneider, A., Stachelhaus, T., and Marahiel, M. A. (1998). Targeted alteration of the substrate specificity of peptide synthetases by rational module swapping. *Mol. Gen. Genet.,* **257,** 308–318.

Schneider, W. P., Nichols, B. P., and Yanofsky, C. (1981). Procedure for production of hybrid genes and proteins and its use in assessing significance of amino acid differences in homologous tryptophan synthetase alpha polypeptides. *Proc. Natl. Acad. Sci. USA,* **78,** 2169–2174.

Schnepp, B., Donaldson, T., Grumbling, G., Ostrowski, S., Schweitzer, R., Shilo, B. Z., and Simcox, A. (1998). EGF domain swap converts a *Drosophila* EGF receptor activator into an inhibitor. *Genes Dev.,* **12,** 908–913.

Schultz, G. E., and Schirmer, R. H. (1979). Principles of Protein Structure. Springer-Verlag, New York.

Schwecke, T., Aparicio, J. F., Molnár, I., König, A., Khaw, L. E., Haydock, S. F., Oliynyk, M., Caffrey, P., Cortés, J., Lester, J. B., Böhn, G. A., Staunton, J., and Leadlay, P. F. (1995). The biosynthetic gene cluster for the polyketide immunosuppressant rapamycin. *Proc. Natl. Acad. Sci. USA,* **92,** 7839–7843.

Serre, V., Guy, H., Penverne, B., Lux, M., Rotgeri, A., Evans, D., and Hervé, G. (1999). Half of *Saccharomyces cerevisiae* carbamoyl phosphate synthetase produces and channels carbamoyl phosphate to the fused aspartate transcarbamoylase domain. *J. Biol. Chem.,* **274,** 23794–23801.

Shao, Z., Zhao, H., Giver, L., and Arnold, F. H. (1998). Random-priming *in vitro* recombination: an effective tool for directed evolution. *Nucleic Acids Res.,* **26,** 681–683.

Short, J. M. (1997). Recombinant approaches for accessing biodiversity. *Nat. Biotechnol.,* **15,** 1322–1323.

Short, J. M. (1999). Method of DNA reassembly by interrupting synthesis. U.S. Patent, no. 5,965,408.

Smith, J., Berg, J. M., and Chandrasegaran, S. (1999). A detailed study of the substrate specificity of a chimeric restriction enzyme. *Nucleic Acids Res.,* **27,** 674–681.

Smithrud, D. B. and Benkovic, S. J. (1997). The state of antibody catalysis. *Curr. Opin. Biotechnol.,* **8,** 459–466.

Sode, K., Watanabe, K., Ito, S., Matsumura, K., and Kikuchi, T. (1995). Thermostable chimeric PQQ glucose dehydrogenase. *FEBS Lett.,* **364,** 325–327.

Souciet, J. L., Nagy, M., Le Gouar, M., Lacroute, F., and Potier, S. (1989). Organization of the yeast URA2 gene: identification of a defective dihydroorotase-like domain in the multifunctional carbamoylphosphate synthetase-aspartate transcarbamylase complex. *Gene,* **79,** 59–70.

Srere, P. A. (1987). Complexes of metabolic enzymes. *Ann. Rev. Biochem.,* **56,** 89–124.

Stachelhaus, T., Schneider, A., and Marahiel, M. A. (1995). Rational design of peptide antibiotics by targeted replacement of bacterial and fungal domains. *Science,* **269,** 69–72.

Starzyk, R. M., Burbaum, J. J., and Schimmel, P. (1989). Insertion of new sequences into the catalytic domain of an enzyme. *Biochemistry,* **28,** 8479–8484.

Stemmer, W. P. C. (1994). DNA shuffling by random fragmentation and reassembly: *In vitro* recombination for molecular evolution. *Proc. Natl. Acad. Sci. USA,* **91,** 10747–10751.

Stenlund, P., and Tibell, L. A. (1999). Chimeras of human extracellular and intracellular superoxide dismutases. Analysis of structure and function of the individual domains. *Protein Eng.,* **12,** 319–325.

Takai, T., Yokoyama, C., Wada, K., and Tanabe, T. (1988). Primary structure of chicken liver acetyl-CoA carboxylase deduced from cDNA sequence. *J. Biol. Chem.,* **263,** 2651–2657.

Tang, J., James, M. N., Hsu, I. N., Jenkins, J. A., and Blundell, T. L. (1978). Structural evidence for gene duplication in the evolution of the acid proteases. *Nature,* **271,** 618–621.

Thoden, J. B., Kappock, T. J., Stubbe, J., and Holden, H. M. (1999). Three-dimensional structure of N(5)-carboxyaminoimidazole ribonucleotide synthetase: A member of the ATP grasp protein superfamily. *Biochemistry,* **38,** 15480–15492.

Thornton, J. M., and Sibanda, B. L. (1983). Amino and carboxy-terminal regions in globular proteins. *J. Mol. Biol.,* **167,** 443–460.

Tsien, R. Y. (1998). The green fluorescent protein. *Ann. Rev. Biochem.,* **67,** 509–544.

Wah, D. A., Hirsch, J. A., Dorner, L. F., Schildkraut, I., and Aggarwal, A. K. (1997). Structure of the multimodular endonuclease *Fok*I bound to DNA. *Nature,* **388,** 97–100.

Wah, D. A., Bitinaite, J., Schildkraut, I., and Aggarwal, A. K. (1998). Structure of *Fok*I has implications for DNA cleavage. *Proc. Natl. Acad. Sci. USA,* **95,** 10564–10569.

Wang, W., Kappock, T. J., Stubbe, J., and Ealick, S. E. (1998). X-ray crystal structure of glycinamide ribonucleotide synthetase from *Escherichia coli. Biochemistry,* **37,** 15647–15662.

Wieligmann, K., Norledge, B., Jaenicke, R., and Mayr, E. M. (1998). Eye lens betaB2-crystallin: circular permutation does not influence the oligomerization state but enhances the conformational stability. *J. Mol. Biol.,* **280,** 721–729.

Wright, G., Basak, A. K., Wieligmann, K., Mayr, E. M., and Slingsby, C. (1998). Circular permutation of βB2-crystallin changes the hierarchy of domain assembly. *Protein Sci.,* **7,** 1280–1285.

Wu, R., Ruben, G., Siegel, B., Jay, E., Spielman P., and Tu, C. D. (1976). Synchronous digestion of SV40 DNA by exonuclease III. *Biochemistry,* **15,** 734–740.

Xu, G. L., and Bestor, T. H. (1997). Cytosine methylation targetted to pre-determined sequences. *Nat. Genet.,* **17,** 376–378.

Xue, Q., Ashley, G., Hutchinson, C. R., and Santi, D. V. (1999). A multiplasmid approach to preparing large libraries of polyketides. *Proc. Natl. Acad. Sci. USA,* **96,** 11740–11745.

Yoshida, H., Kojima, K., Witarto, A. B., and Sode, K. (1999). Engineering a chimeric pyrroloquinoline quinone glucose dehydrogenase: improvement of EDTA tolerance, thermal stability and substrate specificity. *Protein Eng.,* **12,** 63–70.

Zalkin, H., and Nygaard, P. (1999). Biosynthesis of purine nucloetides. In: *"Escherichia coli and Samonella: Cellular and molecular biology"* (F. C. Neidhardt, ed.), pp. 561–579. ASM Press, Washington, D.C.

Zebala, J., and Barany, F. (1991). Mapping catalytically important regions of an enzyme using two-codon insertion mutagenesis: a case study correlating β-lactamase mutants with the three-dimensional structure. *Gene,* **100,** 51–57.

Zhao, H., Giver, L., Shao, Z., and Arnold, F. H. (1998). Molecular evolution by staggered extension process (StEP) *in vitro* recombination. *Nat. Biotechnology,* **16,** 258–261.

This Page Intentionally Left Blank

RATIONAL EVOLUTIONARY DESIGN: THE THEORY OF *In Vitro* PROTEIN EVOLUTION

By CHRISTOPHER A. VOIGT,* STUART KAUFFMAN,† and ZHEN-GANG WANG‡

*Biochemistry Option, Divisions of Biology and Chemistry and Chemical Engineering
‡Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena,
California 91125, and †Bios Group INC, Santa Fe, New Mexico 87501

I. INTRODUCTION

Darwin (1859) argued that species evolve due to the driving force of natural selection, requiring only reproduction, heredity, and new diversity. In general, the experimental protocol of directed evolution does not fundamentally differ from Darwin's original hypothesis. Mutagenesis and recombination are applied to the wild-type DNA sequence, creating a diverse mutant library. In stepwise evolution, the mutant library is screened for the desired function and either the fittest protein becomes the parent to the next generation or several of the fitter mutants are recombined (Moore *et al.,* 1997; Zhao and Arnold, 1997a). In contrast to natural evolution, where the goal of an organism is simply survival and reproduction, directed evolution experiments are controlled by the researcher, who defines the goal as well as the methodology by which it may be achieved. By iterating between mutation, recombination, and screening steps, large improvements in the desired function can be obtained. Alternatively, methods have been developed to evolve populations, such as phage display, pooling algorithms, or selection methods based on coupling the performance of the protein with its genetic material (Roberts and Ja, 1999). For both stepwise and population methods,

0065-3233/00 $35.00

in vitro evolution has proved a valuable tool in the development of enzymes with altered function (e.g., functions such as stability, selectivity, or activity in chemical environments not intended by nature) (Moore and Arnold, 1996; Kuchner and Arnold, 1997; Skandalis *et al.,* 1997).

Although evolution is a superb method to design proteins, it is slow. A process that naturally occurs over time scales of millions of years is infeasible for laboratory work. For directed evolution to be an acceptable approach to protein design, it must have reliable project times in the range of months. This can be achieved by discovering the underlying processes that drive evolution and using them to optimize the experimental parameters. The basis of these driving forces is analogous with the protein-folding problem. A protein sequence rapidly folds into a well-defined structure despite a combinatorial explosion of conformational possibilities (Levinthal, 1969). For a sequence to find its correct folded state, it must either sample all conformations or follow guiding principles that dominate the folding process. These principles lead to the energy landscape paradigm in which the protein conformation is guided by the landscape features until it descends to the minimum energy ground state (Frauenfelder *et al.,* 1991; Dill *et al.,* 1995). Design through evolution can be treated similarly. Here, the moves are not made by conformational changes, but rather through mutations. Again, the space is huge and cannot be searched exhaustively, so an effective search algorithm is required.

There is a wealth of theory on the process of evolution that is largely overlooked in the experimental literature. Part of the difficulty is the abundance of jargon that is specific to relatively small clusters of theoretical literature. Theoretical studies on protein evolution, RNA evolution, DNA evolution, and algorithms in computer science often have interchangeable results. For instance, motifs that have important implications for experiments with proteins emerge from RNA secondary structure studies. However, using the information requires a substantial translation of the language.

In this chapter, we glean the most relevant results from different backgrounds and combine them in a way that is useful to researchers applying evolutionary methods to protein design. We focus on controlling the evolutionary dynamics with an emphasis on relationships between experimental parameters that lead to phase transitions and upper and lower limits. The backbone and motivation behind the theory—information that is necessary to understand the results—are explained in Section II. The implications of the theoretical work, as applied to current methods in directed evolution, are explored in Section III. Finally, results that are important in driving future evolution experiments are described in Section IV. This chapter has been written to appeal to two audiences: the experimentalist, as a tutorial and review, and the theorist, as a manifesto that exposes the desires and weaknesses of the theory of *in vitro* protein evolution.

II. Sequence Space and Fitness Landscapes

The fitness landscape analogy was first described by Sewall Wright. He postulated that by mutating genes on or off there is a change in the ability of the organism to survive (Wright, 1932). From this observation, a multidimensional surface is formed by combinations of on and off genes on which selection drives a population toward a peak, representing an optimal combination of genes. John Maynard Smith extended the genetic basis of a fitness landscape to the molecular world by describing sequence space as a connected network of all protein sequences (Maynard Smith, 1970). Each point in sequence space represents a unique amino acid combination. Because the sequence determines the properties of the protein, there will be a fitness associated with each point. In this context, the fitness is a quantitative, generic ability of the protein to perform its function, or, more fundamentally, it is the properties undergoing selective pressure. The assignment of a fitness to each sequence point produces a fitness landscape on which a evolving protein walks via point mutations, recombination, or more complex moves (Fig. 1, see color insert). The descriptive geological language of this process is beautiful: mountain ranges representing condensed areas of high fitness, small peaks in which a protein can become trapped, ridges connecting ranges, and the elusive Everest, representing the globally fittest protein (Wright, 1967; Kauffman, 1993).

Searching sequence space is not trivial. First, it is hyperastronomically large: For an average protein of 300 residues, the number of sequence points is 20^{300}. In comparison, the most mutants that can be observed through screening or selection is 10^5 to 10^{13}. Several properties of the landscape allow it to be effectively searched despite the small fraction of sequences that can be observed at each step. Although the space is large, it is also highly connected. Each sequence has $(A - 1)N$ neighbors, where N is the number of residues and A is the total number of kinds of amino acids. Distance is measured by the number of mutational steps required to convert one sequence into another, known as the Hamming distance (Hamming, 1950). Any two points are separated by a maximum of N mutational moves, implying that the total number of steps required to travel from a random point to the global maximum is experimentally attainable. However, the number of possible paths between sequences

Legends for Color Insert

FIG. 1. A two-dimensional projection of the hyperdimensional fitness landscape. In this simplified representation, sequence space is shown for a 4-mer where the colors represent amino acid types. The all-blue sequence is the global optimum; the lower fitness peaks are local optima. The problem of *in vitro* evolution is how to search this space effectively, without becoming trapped at a suboptimal fitness.

FIG. 13. The predicted entropy distribution for subtilisin E as determined by a mean-field treatment of the structural model. When all amino acids are equally allowed at a position, $s_i = \ln 20 \approx 3.0$. The red lines are the positions at which mutations discovered by directed evolution improved the thermostability. The blue lines are for mutations that improved the activity (hydrolysis of a peptide substrate) in aqueous dimethylformamide. The bars indicate the average and standard deviation of the structural entropies.

FIG. 19. A cartoon of neutral evolution aiding evolutionary search (top). The periods of adaptive evolution are shown in red and neutral evolution in black. If only an adaptive search was performed, the sequence would quickly become trapped at the first local optimum. Neutral drift allows a population to extend along a fitness ridge, until a path for adaptive evolution is discovered. The bottom graph shows the evolutionary dynamics of a RNA simulation toward a target structure. The population average of the distance to the target is plotted against time. The dots are the projection coordinate of the distribution of structures found in the population at a given time. The flat regions are epochs dominated by neutral drift and continuous transitions. The sudden jumps are adaptive regions dominated by discontinuous transitions. Note that the jumps are also characterized by narrowing of the structure distribution in the population. The bottom figure is reprinted from Huynen *et al.* (1996) with permission. Copyright (1996) National Academy of Sciences, USA.

FIG. 20. Movement through sequence space by continuous structural transitions. Each colored oval represents a structural neutral network. The transfer between neutral networks by continuous structural changes (solid arrows) is more likely than a sudden, discontinuous jump (dotted line).

increases exponentially as the Hamming distance between the start and end points increases, making the discovery of the most efficient path difficult. There are also many regions of sequence space where an evolving sequence may become trapped. For instance, if all single mutations from a parent result in less fit mutants, the parent lies on a fitness peak. When this peak does not correspond to the globally fittest sequence, it is referred to as a local optimum and is a trap for a sequence that is evolving through mutagenesis alone. In addition, there are islands of function that cannot be reached through single amino acid substitutions. Finally, most space is void of function, giving rise to the need for a connected pathway or efficient algorithm that can reliably move through the space.

To study a fitness landscape theoretically, each sequence must also have a defined fitness, so the selective advantage of sequences can be quantitatively compared. The connection between sequence and fitness is made through a fitness function. Unfortunately, the understanding of the relationship between sequence and structure is in its infancy and the more relevant connection between sequence and function is currently unattainable. These realities have forced the construction of simplified models to describe the connections between sequence, structure, and function, in order to study the underlying evolutionary trends. Each model is an attempt to capture the element of biopolymers that drives the aspect of evolution addressed. In Section II.A, we present the physical properties of proteins that lead to defining landscape features. In Section II.B, we compare theoretical constructs that are commonly used to study different problems of evolution. Finally, in Section II.C, we discuss characterizations of the gross structural features of the landscape as a means to identify effective search strategies.

A. *Experimental Motivation for Fitness Landscapes*

The structure of a real biopolymer landscape is highly complex. Recently, researchers have started to understand some of the properties of proteins that lead to specific landscape structures. To simplify the problem, it is desirable to identify which physical characteristic of proteins give rise to dynamic trends in evolution. In this section, we review the experimentally deduced properties of coupling and additivity, with the intention of demonstrating that these properties are correlated and discussing the landscape features that they produce.

1. *Coupling*

Consider the following situation. A professor hires two new students who are each capable of producing wonderful research. It is possible, however, that their personalities conflict and nothing is achieved (possibly due to their close proximity). Of course, it is also possible that their personalities do not conflict and their combined efforts are beneficial. If the new students work together well, additional productivity can be achieved. In an ideal scenario, a lab would be constructed that takes full advantage of talent and minimizes conflict. In terms of protein chemistry, the students in this analogy are amino acids. The coupled energy between conflicting amino acids causes a lower fitness; when there is significant coupling, simple optimization is impossible. The inability to optimize the components of a system because of coupling effects is referred to as frustration.

In the case where the fitness of interest is protein stability, the sources of coupling are relatively understood. If two side chains interact through a hydrogen bond or salt bridge to produce a pairwise energy, the stabilizing energy can be lost by mutating either residue (Wells, 1990). Pairwise coupling can also arise out of steric interactions: two residues close in space may have side chains that are too large to fit in adjoined space, leading to van der Waals conflicts. The pairwise nature of many energy contributions has been used to improve the stability of small proteins through computational methods (Dahiyat and Mayo, 1997).

Although the stability problem lends itself to a pairwise description of interactions, coupling becomes more complicated on the functional level and can manifest itself through long-range, and multiple-residue interactions. For a protein to fold, it is important to have a hydrophobic surface (Dill *et al.,* 1995). This implies that there is a contribution to the stabilization energy made by a collective contribution from multiple hydrophobic core residues. In addition, propagation of electrostatic interactions through the protein can lead to coupling between spatially distant residues (LiCata and Ackers, 1995). Mutations can also cause global and local conformational changes, leading to additional coupling. Mutations sometimes induce shifts in the backbone structure that extend non-isotropically from the mutated residue (Skinner and Terwilliger, 1996). In a typical example, NMR studies of a nuclease showed that a single amino acid change can cause structural perturbations that extend 15–30 Å (Wilde *et al.,* 1988).

Interactions that lead to functional improvement are more difficult to identify. Near the active site, mutations can severely disrupt the function by interrupting an essential arrangement of catalytic residues (Knowles, 1991). This effect is not limited to massive disruptions caused by mutating a catalytic residue; it can occur more subtly, by altering properties important to catalysis, such as vibrational modes or internal electrostatic fields (Skinner and Terwilliger, 1996). Mutations far from the catalytic site influence activity through long-range perturbations for enzymes (Schachman, 1993; Spiller *et al.,* 1999) as well as antibodies (Chien *et al.,* 1989).

2. *Additivity*

During experiments, coupling is observed through non-additive mutational effects. Two mutations are considered additive if the combined change in fitness is equal to the sum of the individual contributions of each mutation. Additivity implies a smooth landscape that can be easily searched via an evolutionary walk. This simplest description of two additive mutations is

$$w_{AB} = w_A + w_B, \tag{1}$$

were w_A and w_B are the change in fitness from mutations A and B, and w_{AB} is the combined change. If two mutations are non-additive, then an additional term has to be added to the right-hand side to balance the equation. The magnitude of this term is the degree of non-additivity.

Non-additivity arises from the simultaneous disruption of coupled residues by multiple mutations. Mutations affecting some functions, such as protein-protein and DNA-protein interactions, tend to be remarkably additive, while others, such as mutations in the catalytic site, tend to be non-additive (Jencks, 1981; Wells, 1990). Mutational studies on T4 lysozyme (Zhang *et al.,* 1995) and catalase I (Trakulnaleamsai *et al.,* 1995) indicate that thermostability is largely additive, and therefore relatively uncoupled. Non-additivity is most commonly observed when the mutated residues are close in space and large, or when chemically disparate side chains are introduced.

LiCata and Ackers report that mutations that are not directly in contact can be non-additive (1995). They find that most mutations exhibit some degree of non-additivity that cannot be explained by short-range disruptions. A structural study of mutants of pNB esterase generated by directed evolution supports this observation (Spiller *et al.,* 1999; see chapter by Orencia, Hanson, and Stevens in this volume). In this case, the influence of a mutation was realized through small backbone shifts, spatially distant from the mutated residue. Non-additivity can result when these perturbed regions overlap (Skinner and Terwilliger, 1996).

As the number of mutated sites increases, so does the probability that perturbed regions will interact, and the probability of observing non-additivity increases (LiCata and Ackers, 1995). During the directed evolution of sublisin E to increase its activity under nonnatural conditions, Chen and Arnold found a triple-mutant whose mutations were additive. Further evolution generated a 10-fold mutant with significant non-additivity (Chen and Arnold, 1993). In another directed evolution study, structural arguments were used to explain the non-additivity of successive mutations in pNB esterase (Spiller *et al.,* 1999). From the X-ray structures of the wild type and its eight-fold mutant, it was shown that a mutation early in the evolutionary trajectory fixed a loop that is flexible in the wild type. Mutations in later generations interacted with this loop. If the order of mutation were reversed, these mutations would have been unable to make the same contacts and would have had different effects on the fitness.

B. Assigning a Fitness

In biology, fitness represents the ability of an organism to survive and reproduce and is a simplified description of a combination of properties (cf. Michod, 1999). In directed evolution, the fitness is determined by the properties of interest to the researcher. Because the calculation of even simple properties from the primary sequence is currently unattainable, a fitness function is required to make the connection from sequence to fitness. The use of a fitness function implies a statistical mechanics approach to evolution, where the ensemble properties of adaptive walks are studied based on the underlying structure of fitness landscapes (Kauffman and Weinberger, 1989). By studying such generic representations, the nuances of the individual proteins are lost. The hope is that the behavior of evolution will rely on the gross statistical features of the landscape (Amitrano *et al.,* 1989). The simplified fitness functions make the direct translation between sequence and fitness. More complex fitness functions have been proposed that have an additional step: sequence to structure to fitness (Shakhnovich, 1994; Dill *et al.,* 1995; Govindarajan and Goldstein, 1997b). These models have been used when it was considered important to capture structural elements in describing an aspect of evolutionary dynamics.

1. Simplified Fitness Functions

The simplest fitness function is the uncoupled form, in which the fitness is the sum of individual contributions from each residue, representing the special case where all mutations are additive (Aita and Husimi, 1996; Saven and Wolynes, 1997). This model is similar to the description of an ideal gas, where the energy contributions of the molecules are uncoupled. The fitness of a sequence F is

$$F = \frac{1}{N}\sum_{i=1}^{N} f_i(\alpha_i), \qquad (2)$$

where f_i is the fitness of residue i in state α_i, and N is the total number of residues. This fitness function creates a single-optimum (''Fujiyama'') landscape on which optimization is simple. Each residue could be mutated independently and the fittest sequence would be the combination of the best amino acid at each position, requiring a maximum of

$(A - 1)N$ steps. The assumption of no coupling is severe, but additional multibody terms can be added, similar to the virial expansion of an ideal gas (Saven and Wolynes, 1997).

There has been extensive research in statistical physics on spin glasses that consider the relationship between coupling and frustration and have been extensively used to model biological problems, including protein folding and evolution (Sherrington and Kirkpatrick, 1975; Anderson, 1983; Bryngelson and Wolynes, 1987; Fischer and Hertz, 1991). In a simple form, pairwise coupling is included between residues

$$F = \frac{1}{N}\sum_{<ij>} f_i(\alpha_i,\alpha_j), \tag{3}$$

where $<ij>$ represents the summation over all nearest-neighbor contacts, and the fitness contribution depends on both the state of residue i, α_i, and the state of residue j, α_j. The coupling between residues causes the fitness landscape to become more rugged, and finding the optimal sequence becomes a more difficult search problem. Spin glasses have often been used as a benchmark to study the effectiveness of search methods, such as genetic algorithms (Prügel-Bennett and Shapiro, 1994).

To include higher-order coupled interactions among residues, the *NK*-model introduces the more complex fitness function

$$F = \frac{1}{N}\sum_{i=1}^{N} f_i(\alpha_1,\alpha_2, \ . \ . \ . \ ,\alpha_{K+1}), \tag{4}$$

where the fitness contribution of residue i not only depends on the state i, but also on K other residues (Kauffman and Levin, 1987; Kauffman, 1993). The parameter K can be roughly interpreted as the average number of interactions between residues. The interactions are either nearest-neighbor or randomly distributed. Because the tertiary structure of the protein brings residues into contact that are not close in primary sequence, randomly distributed, long-range interactions are clearly important (Abkevich *et al.,* 1995). When $K = 0$, Eq. (4) simplifies to the simple additive expression of Eq. (2) and represents a smooth landscape with a single optimum. As K increases, the landscape becomes more rugged and contains more local optima. The advantage of the tunable *NK*-model is that it allows the study of evolutionary parameters that are believed to depend primarily on the ruggedness of the landscape.

In the limit $K = N - 1$, the *NK*-landscape becomes the random energy model (Derrida, 1980; Derrida, 1981). The fitness function is written as

$$F = f_i(\alpha_1,\alpha_2, \ldots, \alpha_N), \tag{5}$$

where a random fitness is assigned to each sequence. Most likely, fitness landscapes are not well represented by the extreme limit of the random energy model (Macken and Perelson, 1991). However, this case captures the experimental observation that a single point mutation can have a large effect on the fitness of a protein. One characteristic of the random energy model is a large number of local optima. As the number of residues increases, the heights of these optima decrease asymptotically to the mean fitness of the landscape, referred to as the ''complexity catastrophy'' (Kauffman, 1993). In the absence of landscape features to follow, the adaptive walk is severely hindered.

Coupling is not the only source of frustration in the fitness landscape. When several fitness functions are combined, local optima can arise, even if each individual function is a fully additive property that can be described by Eq. (2). Husimi and Aita (1998b) studied the effect of mutations on reactions comprising multiple steps. Consider a reaction that requires two steps—binding and catalysis. When partial fitness landscapes that have single optima are combined, frustration occurs when a mutation that aids binding hinders catalysis, creating rugged landscapes with many local optima.

Although the simplified fitness functions capture some element of the real fitness landscape, there are several important differences. Experimentally, it is known that single mutations can create inactive enzymes, either through the disruption of the catalytic residues or due to a loss of structural integrity. In the models described above, every sequence has an assigned fitness. This implies that a random sequence can be optimized because, although it has a low fitness, it still lies on a landscape feature. In contrast, a real random sequence can lie in a space completely devoid of function and therefore has no features to follow, making evolutionary optimization impossible. In addition, the number of interactions between residues (K in the *NK*-model) can vary from site to site, creating a distribution of coupling effects in the protein (Kauffman and Weinberger, 1989). For instance, residues that lie on the protein surface are less coupled than those in the core. Finally, the distribution of coupling energies utilized in the models is usually either a Gaussian or uniform distribution. The true energy distribution describing amino acid substitutions and its effect on the evolutionary dynamics is unclear.

2. *Model Proteins*

Model proteins have been used extensively to study molecular evolution. In this section, we discuss the basic components of a model protein. Specific applications of these studies will be presented in Sections III and IV. When a polypeptide folds into a compact protein structure, its topology is dictated by the primary amino acid sequence. Despite a large research effort in the field of protein structure prediction, there is currently no reliable method to predict structure given the sequence. As an alternative, researchers have borrowed techniques from polymer physics and spin glasses to construct models of proteins that capture some of the essential features of structure necessary to study evoltuion. These models usually do not allow the extension to function. Nonetheless, they share an underlying similarity to the simple fitness functions described in the previous section, allowing the model protein results to be generalized. The literature on this topic is extensive; the aim of this section is only to introduce the concept and show how model proteins fit into the larger subject of fitness landscapes.

A common approach to improving the tractability of the problem is to reduce the conformational space that can be explored by the protein structure by defining a regular array of coordinates to which each residue is constrained to produce a lattice model (Shakhnovich, 1994; Dill *et al.,* 1995; Govindarajan and Goldstein, 1997a; Saven and Wolynes, 1997). Although there are specific differences in the construction of a lattice protein, the form of the fitness function is the same. In general, it is defined as

$$F = \frac{1}{N}\sum_{i<j}^{N} f_{ij}(\alpha_i,\alpha_j)\,\Delta_{ij}, \tag{6}$$

where the notation is the same as for the simple fitness functions, and Δ_{ij} is the contact matrix. If residues 1 and 2 are in contact, $\Delta_{12} = 1$; if not, $\Delta_{12} = 0$. The differences among the various models fall into three categories: the topology encoded in the contact matrix Δ_{ij}, the alphabet α, and the energy distribution f_{ij}.

The contact matrix mimics side chain interactions; covalently linked amino acids are not considered to be in contact. Despite the clear similarity to the spin-glass representation, several important difference exist. In spin glasses, the contacts between residues are taken as either nearest-neighbors or randomly placed, while the contact matrix Δ_{ij} is calculated from either the two-dimensional or three-dimensional topology (Fig. 2). In addition to the dimensionality, the lattice can be maximally compact, defined as a lattice structure that has the maximum number of contacts (Govindarajan and Goldstein, 1997a).

The combination of the energy distribution and the number of states that are allowed at each position constitute the alphabet. For model proteins, the complex dynamics of the side chains are often condensed into the energy term f_{ij}, whose purpose is usually to model protein stability. The simplest alphabet reduces the information contained in

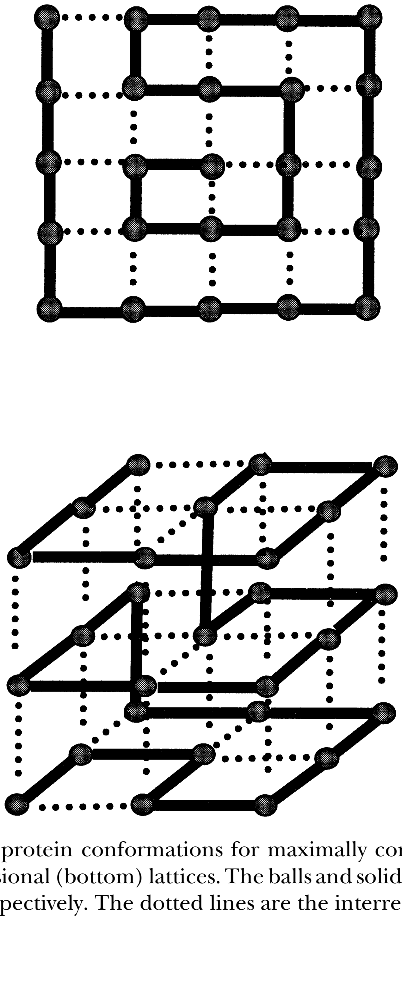


FIG. 2. Examples of protein conformations for maximally compact, two-dimensional (top), and three-dimensional (bottom) lattices. The balls and solid lines represent residues and covalent bonds, respectively. The dotted lines are the interresidue contacts captured by the matrix Δ_{ij}.

the amino-acid side chains to two states: hydrophobic and polar (Dill *et al.,* 1995). This approach is motivated by the observation that protein folding requires a hydrophobic core and hydrophilic surface. The fitness contribution of a hydrophobic/hydrophobic contact f_{HH} is always good, a polar/polar contact f_{PP} is always bad, and a polar/hydrophobic contact f_{PH} is better than or equal to f_{PP} and worse than f_{HH}. The specific fitness contributions vary in the literature. Some problems with a two-letter code have been identified, including highly degenerate ground states (Buchler and Goldstein, 1999a) and an inability to fold into a unique structure (Shakhnovich, 1994). A commonly used alternative is a twenty-letter alphabet constructed by Miyazawa and Jernigan (1985). The fitness contribution f_{ij} is calculated through statistical studies, using X-ray crystal structures, of the probability that two amino acid types are in contact.

3. RNA Secondary Structure

The relationship between sequence space and structure space has been well studied for the RNA secondary structure landscape (Eigen, 1971; Fontana and Shuster, 1987; Eigen *et al.,* 1988; Eigen *et al.,* 1989; Schuster and Stadler, 1994). For RNA, the secondary structure is defined by strong energetic preferences for G≡C and A=U base pairing and is determined by the primary sequence. The secondary structure is composed of loops (unpaired bases connected by a pair), stacks (rows of base pairs), bulges (unpaired regions with more than one branch emanating from them), and tails (unpaired end vertices). The secondary structure is a combination of these elements. The fitness of the structure is either the summation of the free energies of the independent structural components or is obtained through a more tedious calculation involving the kinetic rates of formation and degradation of the structure. In contrast, protein secondary structure, while similarly dependent on the sequence, is not nearly as simply defined and is often the result of less quantifiable, long-range forces. A sharp distinction between similar and dissimilar structures is particularly advantageous for the study of neutrality, which will be presented in Section IV.

In addition to providing a well-defined secondary structure, the dynamics of RNA evolution provide insight into the dynamics of a population on the fitness landscape. One method is to model evolution by using deterministic classical kinetics, which has been described extensively by Eigen and co-workers (Eigen, 1971; Eigen *et al.,* 1988; Eigen *et al.,* 1989). Kinetic equations are introduced to model the evolution of a population of M sequences as in a chemical reactor. The kinetics for the replication of sequence I_i are captured by the following elementary reactions:

$$I_i \xrightarrow{A_iQ_{ii}} 2I_i \tag{7}$$

$$I_i \xrightarrow{A_iQ_{ik}} I_k + I_i \tag{8}$$

$$I_i \xrightarrow{D_i} 0, \tag{9}$$

where A_i is the replication rate constant, D_i is the death rate, and I_k is the mutated sequence that is generated from I_i. The matrix Q_{ik} is the error rate for the transcription of I_k from I_i and can be derived from the mutation rate p_m as

$$Q_{ik} = (1 - p_m)^{N-d_{ik}} p_m^{d_{ik}}, \tag{10}$$

where d_{ik} is the Hamming distance and N is the sequence length. The effective fitness of each sequence is calculated from the relative replication and death rates, $F_{ij} = A_iQ_{ii} - D_i$. The sequence space is similar to proteins, except the alphabet is smaller (four nucleotides versus twenty amino acids), thus reducing the dimensionality.

The distribution of mutant sequences will asymptotically approach the dominant right-hand eigenvector of the fitness matrix F. This stationary distribution defines the quasispecies as a master sequence I_0 and a cloud of frequent mutants that is held fixed by a balance of the opposing forces of mutation and selection. A phase transition based on the mutation rate exists where the population dynamic changes from a localized distribution of related sequences to a randomly drifting cloud of mutants (Fontana and Shuster, 1987; Eigen *et al.,* 1988). When the mutation rate exceeds this transition point (referred to as the error threshold), the population loses the information of the master sequence. The heredity between parent and offspring is broken, making adaptive evolution impossible. An understanding of the error threshold is important in determining the optimal mutation rate (Section III.A.3) as well as the behavior of a population on a neutral network (Section IV.B.3).

C. Characteristics of Fitness Landscapes

In this section, we describe several statistical quantities that capture the gross structural features of the landscape. In Section II.C.1, we discuss the experimental observations of the tolerance of proteins to mutation and describe a correlation between the tolerance and the interresidue coupling. In Section II.C.2, landscape ruggedness is described in terms of the fitness correlation between the parent and mutant sequences. Finally, in Section II.C.3, we describe landscape isotropy as a means to define a local region of sequence space that can be statistically modeled with data generated from evolution experiments.

1. *Tolerance*

Proteins are surprisingly resilient toward mutation. Tolerance is defined as the ability of a protein to undergo mutation without disrupting its fitness or structure. Within a protein, there is a distribution of tolerances. Some sites that are essential for function may not accept any mutations, while other positions can accept almost all amino acid substitutions with little effect. We can speak of both structural and functional tolerance, depending on the property that is robust to mutation. There is some connection between structural and functional tolerance because mutations that tend to be bad for structure also are bad for function. It is not expected that the overlap between structure and function is exact, however, since structural perturbations can conceivably retain function and, conversely, mutations that do not effect the structure could destroy function [such as mutations of essential catalytic residues (Hellinga *et al.,* 1992; Shoichet *et al.,* 1995; Suzuki *et al.,* 1996b)].

Tolerance was beautifully illustrated experimentally by Reidhaar-Olson and Sauer (1988, 1990), who explored the effects of all point mutations on a helix of λ repressor. Some positions allowed a surprising number of substitutions while retaining a measurable degree of DNA binding. In a similar study, Brown and Sauer characterized the tolerance of Arc repressor by measuring the effects of substituting alanine at multiple positions in the sequence (Brown and Sauer, 1999). Surprisingly, alanine mutations were tolerated at twenty-two residues (fifteen simultaneously, out of a total fifty-two residues) without severely disrupting the structure.

Functional tolerance is a significant factor for the success of directed evoluiton. A protein that is functionally tolerant allows many mutations without disrupting the fitness, making it more likely that there is a connected path in sequence space of single mutations that leads to regions of higher fitness. Tolerance also affects the quality of the mutant library. If the protein is functionally intolerant, the mutant library will consist of mostly inactive proteins.

Structural tolerance is also a crucial property for the success of directed evolution (Voigt *et al.,* 2000b). Maintaining structure is required for the acquisition or fine-tuning of any other property (Hawrani *et al.,* 1994). For example, it has been suggested that properties such as stability and activity are coupled (Shoichet *et al.,* 1995). A mutation that is good for stability is bad for activity and a mutation that is good for activity is bad for stability. Directed evolution experiments have shown that many enzymes have not reached a point in evolution at which this coupling is on a fundamental physical level (Zhao and Arnold, 1999). The reason there is apparent coupling between the properties is simply that there is a small probability of improving either. Consider P_S, the probability of not destabilizing the structure through mutations, and P_A, the probability of improving the activity. The probability of improving activity without destabilizing the structure is then $P_{A+S} = P_A \times P_S$. Typically, the individual probabilities are small, so P_{A+S} is very small. The effect of structural tolerance is to increase P_S, thereby increasing the combined probability P_{A+S}. When a protein exhibits structural tolerance, functional space can be more thoroughly explored.

Tolerance is related to the properties of coupling and additivity. Sites that are weakly coupled (additive) tend also to be tolerant, such as residues that lie on the surface (Brown and Sauer, 1999). There are weakly and strongly coupled residues in the protein sequence and, similarly, there is a distribution of tolerance effects (Reidhaar-Olson and Sauer, 1988; Brown and Sauer, 1999). A correlation was noted between the solvent accessibility of a residue and its tolerance, indicating that coupling interactions are important in determining tolerance (Hawrani *et al.,* 1994; Goldman *et al.,* 1998). When a residue's side chain extends out from the surface, it is involved in fewer coupling interactions and thus more substitutions can be performed without disrupting the structure. Model proteins have been extensively used to quantitatively define the relationship between coupling and tolerance. Using a cubic lattice model, Wolynes and co-workers demonstrated that the residues with the most contacts tend to be the least tolerant (Saven and Wolynes, 1997).

2. *Correlation*

The correlation of the landscape measures the fitness similarity between a sequence and its *d*-mutant neighbors, where *d* is the number of mutations. As a sequence accumulates mutations, the fitness is increasingly altered. On smooth landscapes, the rate of fitness change is slow and therefore the landscape is correlated. Conversely, on rugged landscapes, the rate is more rapid and the landscape is uncorrelated. Studies of the relationship between fitness and distance have been used to quantitate the correlation among population ensembles on the RNA landscape (Fontana and Shuster, 1987), recombination dynamics (Bornholdt, 1998), and the success of genetic algorithms (Manderick *et al.,* 1991; Jones and Forrest, 1995).

Weinberger (1990) developed the correlation function as a means to measure the randomness of a fitness landscape statistically. The correlation function is derived under the assumption that the landscape is statistically isotropic (see Section II.C.3) and is written as

$$R(d) = \frac{\langle F_c F_{c+d} \rangle - \langle F_c \rangle \langle F_{c+d} \rangle}{\langle F_c^2 \rangle - \langle F_c \rangle^2}, \tag{11}$$

where d is the distance in sequence space from the reference point c (in Hamming units), and F is the fitness. The denominator of Eq. (11) normalizes $R(d)$. The correlation function of a landscape is determined through a random walk, in which the fitness is recorded as a function of the distance from the starting point (Fig. 3). The correlation function has been shown to decay smoothly for most landscapes (Stadler, 1992; Schuster and Stadler, 1994; Stadler, 1996). In the limit of the random

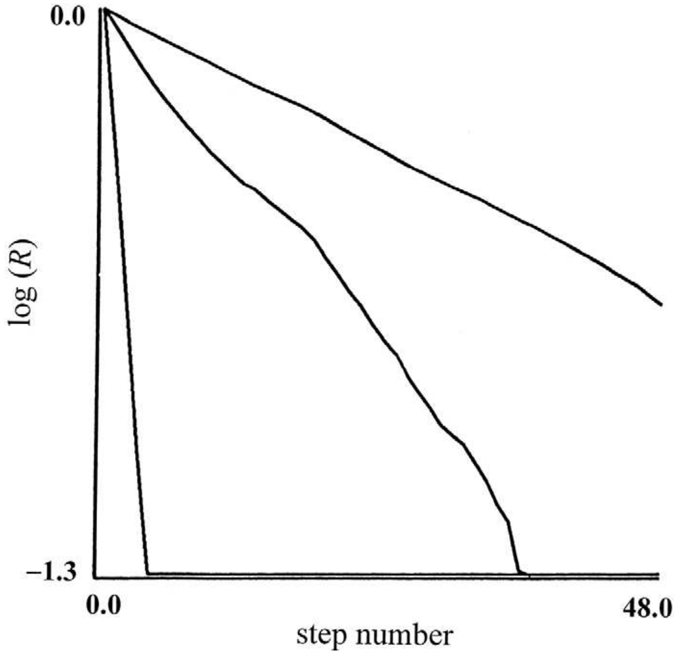


Fig. 3. Semi-log plot of the correlation function for random walks on the *NK*-landscape for $N = 96$ and (from top to bottom) $K = 2$, 8, and 48. In calculating the correlation function, neighbors of each residue are chosen randomly with equal probability. Reprinted from Weinberger (1990) with permission, © 1990 by Springer-Verlag.

energy model, $R(d)$ decays very sharply as a delta function because, by definition, there is no correlation in the fitness between sequence points. The decay of the fitness function is a means to determine a statistical estimate for the ruggedness of the fitness landscape.

It is of interest to condense the information held in the correlation function to a single value. For this purpose, the distance beyond which two points are essentially uncorrelated is defined as the correlation length $1/\tau$ (Macken and Stadler, 1993). On more rugged landscapes, the decay of $R(d)$ with d will be rapid, corresponding to a small correlation length. The converse is true for correlated landscapes. The correlation length is also related to the tolerance of a protein. A small correlation length indicates that single mutations will have large effects on the fitness; a longer correlation length implies the opposite. Therefore, a long correlation length indicates that the protein is tolerant to mutation.

3. *Isotropy*

The effect of having many peaks, ridges, and plateaus in the fitness landscape is that the features of the landscape depend on where the viewer is standing, referred to as the isotropy of the landscape (Stadler and Grüner, 1993; Macken and Stadler, 1993; Shuster and Stadler, 1994; García-Pelayo and Stadler, 1997). Using the geology of the United States as an analogy, the Midwest would be isotropic because no matter where you are standing it looks the same. Isotropy does not necessary imply flatness; the Appalachian Mountains are also isotropic because everything is green and mountainous. In contrast, the Southwest, which contains deserts plains, plateaus, and mountains, would be anisotropic. More rigorously defined, a landscape is isotropic when the statistics of the landscape are the same when referenced from any sequence point. Consider starting at a random point in the landscape and then taking steps in random directions. For an isotropic landscape, the list of fitnesses that results from this process will not depend on the starting position (Weinberger, 1990). One interpretation of anisotropy is the existence of many correlation lengths in a single landscape. A formal definition of anisotropy, based on graph theory, has been derived by Stadler and Grüner (1993).

Most of the simple fitness functions, including the spin-glass models and the *NK*-model, are isotropic (Schuster and Stadler, 1994). However, it is expected that real sequence spaces are highly anisotropic with most regions devoid of fitness, and functional regions that are rich with landscape features. A real protein landscape can appear isotropic in two ways. The first is if a subset of the space is defined—for instance, only the functional sequences are considered. Further, the landscape can appear more isotropic as the length scale is reduced (Flyvbjerg and Lautrup, 1992; Macken and Stadler, 1993). If only a few mutations have accumulated, the landscape may appear isotropic, but as the distance from the starting sequence is increased, the landscape becomes more anisotropic. The transition between the observation of local and global features occurs at a distance from the reference equal to the correlation length of the landscape (Schuster and Stadler, 1994). The degree of isotropy in the fitness landscape is thus relevant in analyzing the screening data generated by an evolution experiment. It is important to understand the anisotropy of the landscape in order to determine whether the data are representative of global features or only the local landscape.

A local description of the fitness landscape may be sufficient in guiding evolution experiments. The range of sequence space that can be defined as local is dependent on the correlation and isotropy of the landscape. The advantage of viewing the landscape features from a local perspective is that it allows a single feature of the landscape to be studied without the concern of scaling the results to a more global description. Urabe and co-workers propose that the structure of the local landscape can be estimated by defining a *local* smoothness s to characterize the region of the fitness landscape that is correlated and isotropic:

$$s \equiv \frac{|<\Delta F>|}{(<\Delta F^2> - <\Delta F>^2)^{0.5}}, \tag{12}$$

where $<\Delta F>$ is the average change in fitness of all single mutants of the parent enzyme (Trakulnaleamsai *et al.,* 1995). The property of local smoothness is unrelated to the global smoothness described previously. An uncoupled landscape can be locally rugged and a random energy landscape can show local smoothness. If mutations are primarily additive, then an optimized property increases smoothly as the number of beneficial mutations increases (Aita and Husimi, 1996). However, as more mutations are accumulated, the sequence moves further in sequence space from the point at which the smoothness was calculated. As the landscape becomes more rugged, the smoothness measured on the original sequence will give less information about the landscape.

III. Modeling Directed Evolution

In this section, we explore the theoretical basis for each process in directed evolution: mutation, recombination, and screening. Although this section is subdivided to discuss each process separately, the evolutionary parameters cannot be treated entirely independently. For example, the limited number of mutants that can be screened imposes restrictions on the optimization of other parameters. For all the processes, the optimum parameters depend on the structure of the fitness landscape. Biological landscapes are likely to be complex and filled with many features that are difficult to navigate. To reach fitness peaks without being constrained by local optima or islands, it is necessary to develop methods that allow evolution to make moves larger than point mutagenesis, such as sexual recombination (Stemmer, 1994a), or using higher mutagenesis rates on smaller regions of the gene (Dube *et al.,* 1993; Miyazaki and Arnold, 1999). This section explores the theoretical background of these methods to provide a basis for the optimization of experiments.

The implementation of cycles of mutation, recombination, and selection also describes the genetic algorithm, an evolutionary optimization technique in computer science (Holland, 1975; Mitchell, 1998). Although observations of nature initially inspired genetic algorithms, much of the subsequent research has lost its relevance to biology. However, some of the results from computer science can be applied to modeling processes in directed evolution, including appropriate mutation and recombination rates, the optimal recombination parameters, and the appropriate mutant library size.

There are several essential differences between genetic algorithms and directed evolution. Genetic algorithms often incorporate populations of solutions, rather than allowing a single sequence to continue to the next generation, as is the design of most stepwise evolution experiments. In addition, results in computer science are presented without concern for the sampling restrictions on the mutant library because the time required to screen a mutant on a computer is fast compared to experiments. In a related problem, the methods introduced to probe the structure of the fitness landscape require large data sets that are not experimentally attainable. The technique to determine the optimal parameters has to be experimentally feasible, preferably using data that are already generated by the experiment. Finally, in protein evolution, mutations are made on the DNA level and then transcribed to amino acid changes. The structure of the genetic code will affect the behavior of an evolutionary walk.

Because the landscape of real proteins is unknown, most of the results we describe in Section III rely on assumptions discussed in Section II. The results are presented for a range of different theoretical landscapes—for example, the random energy model and the uncoupled case—with the assumptions that the real protein landscape lies between these bounds and can be described statistically. Determining the most effective combination of parameters, adjusting them according to the landscape features, and demonstrating how they change as a sequence evolves characterize a proactive evolutionary strategy.

A. Mutation and Selection as an Adaptive Walk

It is convenient to consider evolution as an adaptive walk on the fitness landscape. Driven only by mutation and selection, an adaptive walk is a powerful search technique in computer science as well as *in vitro* evolution (Brady, 1985; Borstnik *et al.,* 1987; Montoya and Dubois, 1993). The speed and sampling ability of the adaptive walk determine the effectiveness of the evolution experiment. In modeling the adaptive walk, it is important to understand how mutations are applied experimentally. Random point-mutagenesis methods make mutations in single bases on the DNA level. Generally, mutations are applied through error-prone PCR by altering the manganese ions in the reaction mixture, allowing mutation rates ranging from 0.11% to 3% (Arnold and Wintrode, 1999; Daugherty *et al.,* 2000). In experimental as well as theoretical studies, it is often assumed that the mutations are made based on a Poisson distribution. In reality, because of biases in transcription, the underlying distribution of mutations is unknown and may vary significantly from this assumption. In Sections III.A.1 and III.A.2, we discuss the behavior of adaptive walks based on the sampling ability of the mutant library and the ruggedness of the fitness landscape. In Sections III.A.3 and III.A.4, we examine the optimal mutation rate based on the size of the screening library, the landscape ruggedness, and the position of the sequence on the fitness landscape.

1. Adaptive Walks with Limited Sampling

The behavior of the adaptive walk is determined by the relationship between the mutation rate and the number of mutants that can be screened. An adaptive walk occurs when an uphill step in a random direction is chosen for each generation. A steepest ascent walk is slightly different: the walk is taken in the direction of the largest fitness improvement, making the process more deterministic than stochastic (Schober *et al.,* 1993). Directed evolution can proceed by either method, as determined by the relative size of the number of mutants screened with respect to the mutation rate. If the screening library is large enough to sample all possible permutations of the average number of mutations, then the evolutionary path is likely to follow a steepest ascent walk; otherwise, it behaves more like an adaptive walk. This is important when determining the experimental reproducibility. Given the same initial sequence, a steepest ascent walk will always converge to the same optimum, while an adaptive walk can take different paths. Whether the optimal search method is adaptive or steepest ascent will depend on the cost of screening mutants, noise in the fitness measurements, and the reduced sequence space dimensionality caused by the genetic code.

Husimi and Aita (1998a) studied the effect of the screening library size M created by d-fold mutations on the ability of the walker to reach the global optimum in an uncoupled landscape. They examined three types of search strategies based on the size of the screening library M with respect to the total number of d-fold mutants M_d and defined a lower limit to the number of mutants screened: $\log_{10} M_c \approx 0.8d \sim 1.5d$. For a 300-residue protein and a mutation rate of two per sequence, $M_d = 3.2 \times 10^7$ and $M_c \cong 1000$. The search strategies are ''shotgun'' for $M \leq M_c$, ''carpet-bombing'' for $M \geq M_d$, and ''intermediate'' for $M_c < M < M_d$. The carpet-bombing search is near deterministic and guaranteed to find the global optimum in the additive landscape. However, because d must be small so that all mutants can be evaluated, the method requires more generations. Shotgun searches have the advantage of evolving more rapidly than carpet-bombing searches. A walker taking the shotgun search strategy has the potential of being trapped in a local optimum on landscapes that are more rugged than the uncoupled assumption. The optimal strategy depends on the cost of screening mutants. If the cost is high, then the walk should tend towards the shotgun type; if it is low, it should tend towards the carpet-bombing type.

Noise in the experimental measurements can change a walk from being steepest ascent to adaptive by requiring a larger critical number of sampled mutants. Experimentally, there is significant error in determining the fitness of an individual mutant, leading to the possibility of false positives continuing to the next generation and false negatives being overlooked by the screen. Levitan and Kauffman (1995) have studied the effect of experimental noise on random uphill walks on NK-model landscapes. Noise increasingly hinders the adaptive walk as the sequence becomes more optimized. At some critical height, the walker cannot continue to climb and hovers at this height. The height of the critical point decreases as the noise increases, causing the walker to ''melt'' off the fitness peak. An increase in noise also makes the walk slower as the walk wastes time by accepting false positives and rejecting false negatives. The effect of a large mutation rate and a small selection strength is similar to that of noise in the measurements (Wagner and Gabriel, 1990). The population will fall to a position in the fitness landscape that is determined by the selection strength and mutation rate.

In evolution, mutations are made on the DNA level, not the protein level. Due to the connectivity of the genetic code, not all amino acid substitutions can be made from a single point mutation. This effectively decreases the dimensionality of sequence space from 19*N* to an average of 5.7*N* by decreasing the number of mutational paths that connect sequence points. This has been incorporated into the *NK*-model to study the maturation of the immune response (Kauffman and Weinberger, 1989). The genetic code is randomly generated, including stop codons (if this mutation is made, the fitness is zero) and synonymous mutations. Because the code reduces the number of available single-point mutations, it affects the adaptive walk by increasing the walk length before a local optimum is reached and decreasing the number of fitter neighbors at each step. Suboptimal mutants are chosen at each generation, making the walk appear more random and requiring that a larger number of mutants be screened.

2. *The Influence of the Fitness Landscape on Adaptive Walks*

The properties of an evolutionary walk change drastically as the landscape becomes more rugged (Kauffman and Weinberger, 1989). The frequency of local optima increases with the ruggedness, and the number of paths that lead to each optimum decreases (Fig. 4). This decreases the probability that a random starting sequence can reach a given optimum through a single-step adaptive walk. A different problem is associated with smoother landscapes. In the *NK*-model, for low *K*, the local optima tend to exist close together and the highest set of peaks is surrounded by large ‘‘drainage basins.’’ These basins create islands of higher fitness that are difficult to reach from an area of low fitness through point mutations alone.

An estimate of the landscape ruggedness can be determined by measuring the path length before an optimum is reached (Macken *et al.*, 1991). The average number of steps that are made during an adaptive climb is a measure of the ruggedness of the landscape. In the uncoupled model, which has only a single optimum, the average number of steps is $N/2$ where N is the number of residues (Kauffman, 1993). As the landscape becomes more rugged, the number of local optima increases, hence making it more likely that a randomly evolving sequence will become trapped. For rugged landscapes, as N increases, the number of local optima increases dramatically.

The fitness landscape also affects the reproducibility of the experiment. If the landscape is very smooth, then there are many possible paths between two sequences (Wang *et al.*, 1996). However, as the ruggedness of the landscape increases, the number of paths sharply decreases.

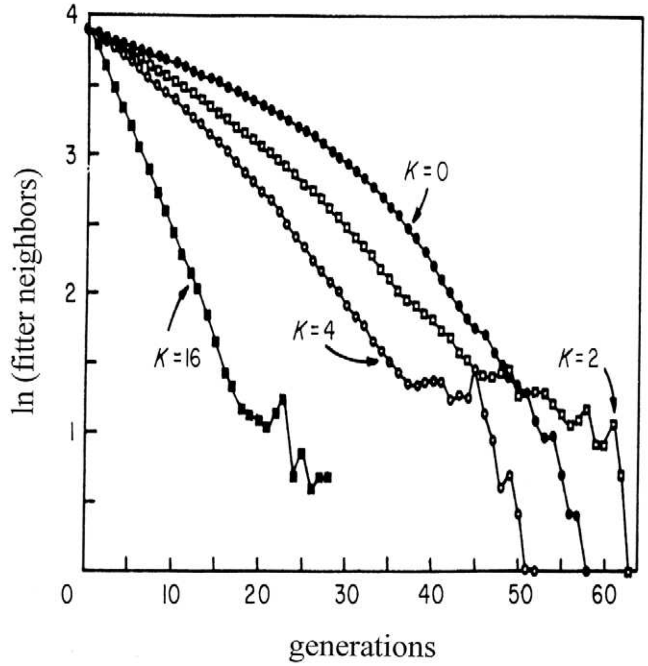


FIG. 4. Logarithm of the mean number of fitter one-mutant neighbors at each adaptive step, plotted against the adaptive step, or generation. Data are taken using the *NK*-model with $N = 96$. As K increases, the rate of fall in the fraction of fitter one-mutant neighbors increases as well. Reprinted from Kauffman and Weinberger (1989) with permission, © 1990 by Academic Press.

Therefore, the walk length before a local optimum is reached will be longer for an adaptive walk than a steepest ascent walk. The walk length R for an adaptive and steepest ascent walk has been estimated using the *NK*-model:

$$R_{adapt} \approx \left(\frac{N}{K+1}\right)\ln[(A-1)(K+1)] \tag{12a}$$

and

$$R_{sa} \approx \left(\frac{A-1}{A\ln A}\right)\left(\frac{N}{K+1}\right)\ln[(A-1)(K+1)], \tag{13}$$

where A is the number of types of amino acids (Weinberger, 1991). Note that the predicted difference in walk lengths between steepest ascent and adaptive walks is only dependent on A. As the ruggedness increases (by increasing K), the walk lengths become shorter. As the size of sequence space becomes large, the length of the adaptive walk grows significantly less rapidly. Equations (12a) and (13) predict upper and lower bounds for the expected number of generations required by *in vitro* evolution to reach an optimum.

Kauffman and Weinberger (1989) studied the length of random walks on NK landscapes and compared the results to the maturation of the immune response. The starting point of the antibody on the landscape was estimated by taking the fraction of receptors on the B-cell that initially bound the substrate (estimated at $1/10^6$) to be the height of the initial sequence in the fitness landscape. They compared the behavior of walks on the NK landscape with that observed experimentally during the accumulation of six to eight somatic mutations. By fitting the distribution of somatic mutations with the predicted distribution, an approximate value of $K \cong 40$ ($N = 112$) was determined. Roughly interpreted, this implies that each residue is coupled to forty other amino acids in the variable region of the antibody. This predicts a very rugged landscape and is a result of considering mutations only in the V-region, analogous to a catalytic site.

In a similar study, protein evolution has been analyzed using the random energy model (Macken and Perelson, 1989). Macken and Perelson calculate the probability of a random walk taking k steps to a local optimum as

$$p(k) = \frac{1}{(k-1)!}\int_0^1 \left(\sum_{i=1}^{D-1}\frac{f_L^i}{i}\right)^{k-1} f_L^{D-1}\, df_L, \tag{14}$$

where $D = AN$ is the number of paths from each point in sequence space and f_L is the fitness of the local optimum. This distribution is approximately constant for $D > 1000$, which is the case for the majority of enzymes, and predicts that the mean number of steps to a local optimum is less than ten, roughly agreeing with Kauffman's estimate from the antibody data.

The benefit of allowing mutants that are less fit than the parent is dependent on the ruggedness of the fitness landscape. If the landscape is very rugged, then an initially bad mutation may be required to ultimately benefit from a coupling interaction. In nature, many mechanisms allow unfit genes to survive. For example, some genes are recessive or are duplicated, allowing the organism to either multiply despite deleterious mutations or hitchhike with fitter genes (Li and Graur, 1991). In computer search algorithms, the benefit of allowing less fit mutants to survive has been well established (Brady, 1985; Amitrano *et al.,* 1989; Pàl, 1995). For instance, a Monte Carlo random walk benefits by including a small temperature term in the reject probability (Metropolis *et al.,* 1953). A very small amount of noise has been shown to improve the outcome of adaptive walks on *NK* landscapes (Levitan and Kauffman, 1995). This is attributed to the evasion of local optima through the ability to surmount energy barriers. Because there are fewer coupling interactions, smoother landscapes benefit less from the inclusion of noise. Practically, the limitations of screening make it difficult to accept downhill steps with the hope of finding more-fit sequences later in the evolutionary path (Moore *et al.,* 1997).

3. *Optimal Mutation Rate*

The optimal mutation rate maximizes the speed of the adaptive walk and is similarly influenced by the number of mutants that can be screened and the structure of the fitness landscape. The general rule has been to use a mutation rate for which the permutations can be effectively sampled during screening. For instance, an average mutation rate of two amino acid substitutions per sequence is only appropriate if the mutant library is large enough to sample most double mutations, based on the assumption that most mutations are deleterious. The structure of the fitness landscape influences the balance between the optimal mutation rate and library size. Certain fitness landscape features are amenable to mutations, so fewer mutants must be screened to achieve the benefit of a higher mutation rate. In the case previously outlined, it would require a smaller library than all double mutant permutations to benefit from a mutation rate of two per sequence. In this section, we explore how the mutation rate is affected by the sampling size and the structure of the fitness landscape, for both population and stepwise *in vitro* evolution.

The quasispecies model defines an optimal mutation rate for evolving populations (Eigen *et al.,* 1988). At the critical mutation rate $p_{m,crit}$ (referred to as the error threshold), the distribution becomes too broad for selection to withstand the dispersion and it wanders stochastically on the fitness landscape. The optimal mutation rate for evolvability should be as close to $p_{m,crit}$ as possible without exceeding it. Indeed, it was found that viral mutation rates are very close to $p_{m,crit}$. By assuming that the mutation probability is the same at each residue, the error threshold in terms of mutation rate $p_{m,crit}$ was derived as

$$p_{m,crit} = \frac{\ln \sigma_0}{N}, \tag{15}$$

where N is the length of the sequence and σ_0 is the fitness superiority of the master sequence I_0. The critical threshold is $\sim O(1/N)$, implying that the optimal average number of mutations per sequence is ln σ_0. The superiority σ_0 is the fitness advantage of the master sequence over its mutant neighbors. If σ_0 is large, then the difference is greater and therefore the error threshold is larger. This is caused by the master sequence's role as an attractor for the mutant distribution. If this attractor is strong (σ_0 is large), then it can balance a large drifting force (mutation rate). Using a perturbation analysis, Eigen *et al.* (1998) evaluated σ_0 for several landscapes in which the long-range correlation between mutants is small, and found that ln σ_0 is typically of the order of three to five.

For quasispecies, the error threshold is lower for finite populations because fitness information can be lost through fluctuations in the population as well as a high mutation rate (Nowak and Shuster, 1989; Bonnaz and Koch, 1998). Under the limit of a finite population M, the error threshold has to be modified

$$p_{m,crit}\ (M) = p_{m,crit}\left(1 - \frac{C}{\sqrt{M}}\right), \tag{16}$$

where $p_{m,crit}$ is the error threshold for an infinite population and C is an empirical constant specific to the fitness landscape. This predicts that the optimal mutation rate decreases as a function of the inverse square root of the population size. Thus, the optimal mutation rate decreases as the sampling ability of the mutant library decreases.

Van Nimwegen and Crutchfield (1999a) derived an expression for the number of mutants screened before reaching the global optimum based on the mutation rate and population size on a "royal road" landscape that is commonly used to study genetic algorithms. The optimal mutation rate is defined as minimizing the number of fitness evaluations E that were required to reach the global optimum. A high mutation rate increases the rate of diffusion and therefore decreases the time required to find improvement. However, a high mutation rate also increases the number of mutants that die by falling off the fitness peak held by the parent. The optimal mutation rate is a balance between these effects. They found that the minimum E associated with the optimal mutation rate is shallow, with steep increases for very small and large mutation rates (Fig. 5). The lower bound is independent of the population size, and the upper bound decreases as the population size decreases. This theory predicts that there is a large region in parameter space in which the evolutionary search performs well and, once in this region, it is not as important to have the exact optimal parameters.

Kauffman and Levin (1987) introduced the concept of long-jump mutagenesis, in which the optimal number of mutations exceeds the correlation length of the landscape. If the walker has the ability to jump beyond the correlation length, the new point is essentially independent of the starting position, making it possible to discover new fitness ranges rather than remaining stuck on the same set of peaks. Multiple mutations can improve the performance of genetic algorithms for smooth landscapes by taking advantage of the correlations between sequence points on the landscape (Manderick *et al.,* 1991). Because landscapes that are more rugged have smaller correlation lengths, a smaller mutation rate is allowed. In the limit of the random energy model, the long-jump benefit is satisfied by single mutations. The experimental concern about using a higher mutation rate is that the density of function drops rapidly as the jump becomes longer, thereby making the probability of discovering improvements negligible. The practical effect of the long-jump theory is to decrease the necessary sampling size for a given mutation rate, based on the correlation length of the landscape.

In quasispecies theory, surpassing the error threshold for small periods provides a beneficial jump in sequence space. Slightly beyond the error threshold, the standard deviation of Hamming distances in the populations is the largest, indicating a spreading out of the population in sequence space, thus allowing for the discovery of multiple mutants that would otherwise be left undiscovered by evolution (Eigen *et al.,* 1988). Bonhoeffer and Stadler (1993) derived the error threshold for complex landscapes as

$$p_{m,crit} \approx \left(\frac{1}{\tau}\right)\frac{K}{N} \tag{17}$$

where N is the sequence length and K is a constant dependent on several landscape-specific factors. The term $1/\tau$ is the correlation length of the fitness landscape. A large $1/\tau$ is indicative of a smooth landscape, and thus Eq. (17) predicts the optimal mutation rate is high. Conversely, a small $1/\tau$ implies a rough landscape and a smaller mutation rate is required, a result consistent with the long-jump hypothesis.

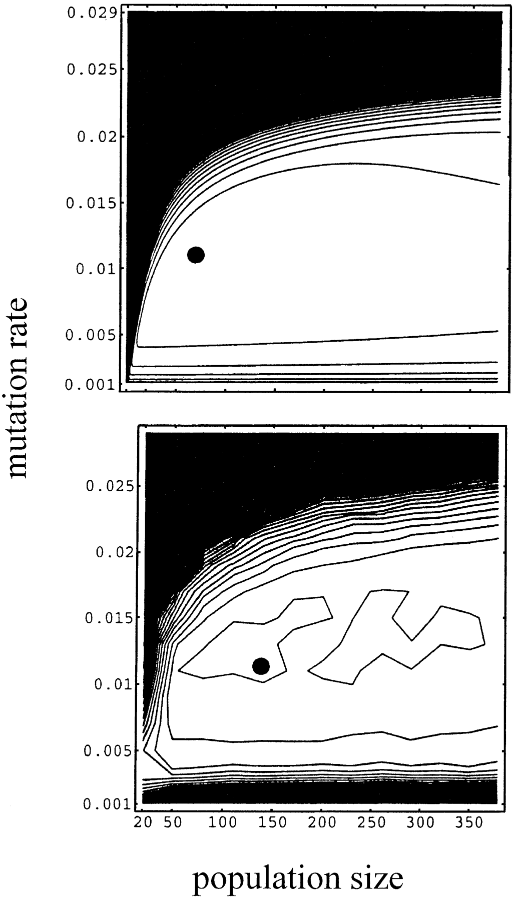


FIG. 5. Contour plots of the search effort surface $E(p_m, M)$, where E is the number of trials required to reach the optimum, p_m is the mutation rate, and M is the population size. Results are shown for the theory presented by Crutchfield and van Nimwegen (upper) and simulations (lower). The x-axis labels are the same for both graphs. The sequence length used is $N = 40$. The dot represents the optimal parameters determined by both methods. Note the very sharp increase in E when suboptimal parameters are used. Within the acceptable parameter settings, the region is relatively flat, suggesting that the specific optimal settings are robust. Reprinted from van Nimwegen and Crutchfield (1999a) with permission.

In directed evolution, the improvement of certain properties, for example, thermostability, is relatively easy. The problem is that these improvements often come at the cost of another desired trait, such as catalytic activity. Using a multistep reaction model, Husimi and Aita

(1998b) found that the walker can escape local optima to a higher fitness at Hamming distance *L,* where *L* is the number of combined fitness landscapes. Therefore, any walker on a multi-fitness landscape can reach the global optimum by repetition of *L*-point mutations. This is similar to the long-jump mutagenesis strategy suggested by Kauffman; correlation of the multistep landscape is dependent on *L* rather than *K* in the *NK* landscape. By making moves that are larger than the correlation length (larger than *L*), more space can be explored.

The desire to have a large mutation rate when there are competing properties to be optimized is countered by the finite number of mutants that can be sampled (Voigt *et al.,* 2000a). This is due to a combination of two effects. First, a higher mutation rate dramatically increases the probability of a mutation to a stop codon. The average fraction of mutated sequences that contain at least one stop codon f_{stop} can be estimated as

$$f_{stop} = 1 - \exp\left(-\frac{3}{63}\langle m\rangle\right), \tag{18}$$

where $<m>$ is the average number of DNA mutations per sequence. When $<m> = 5, f_{stop} = 0.21$ and when $<m> = 20, f_{stop} = 0.61$, indicating that even a moderate mutation rate can cause a large fraction of offspring sequences to contain stop codons, effectively reducing the size of the library. Second, the wild-type sequence is already at a highly optimized fitness and, therefore, most mutations are deleterious. The probability of observing improvement is small for a single mutant and decreases rapidly as the number of mutations increases. Rugged landscapes are more affected by the latter of these effects (Fig. 6). For the mutation of a highly coupled residue to show fitness improvement, it is not only necessary to optimize the uncoupled fitness contribution, but also all the higher-order contributions that are coupled to that residue. It is harder to optimize a coupled rather than uncoupled residue because the probability of optimizing each term is small, causing the optimized mutation rate to decrease as the landscape ruggedness increases.

Because an anisotropic landscape contains many correlation lengths, long-jump theory can be applied to different portions of the sequence (Voigt *et al.,* 2000b). A highly coupled region (such as the catalytic site) has a small correlation length, thus a smaller mutation rate is allowed with a limited mutant library. A highly coupled group of residues requires several simultaneous mutations to demonstrate improvement. Ideally, the optimal mutation rate equals that of the maximum number of resi-

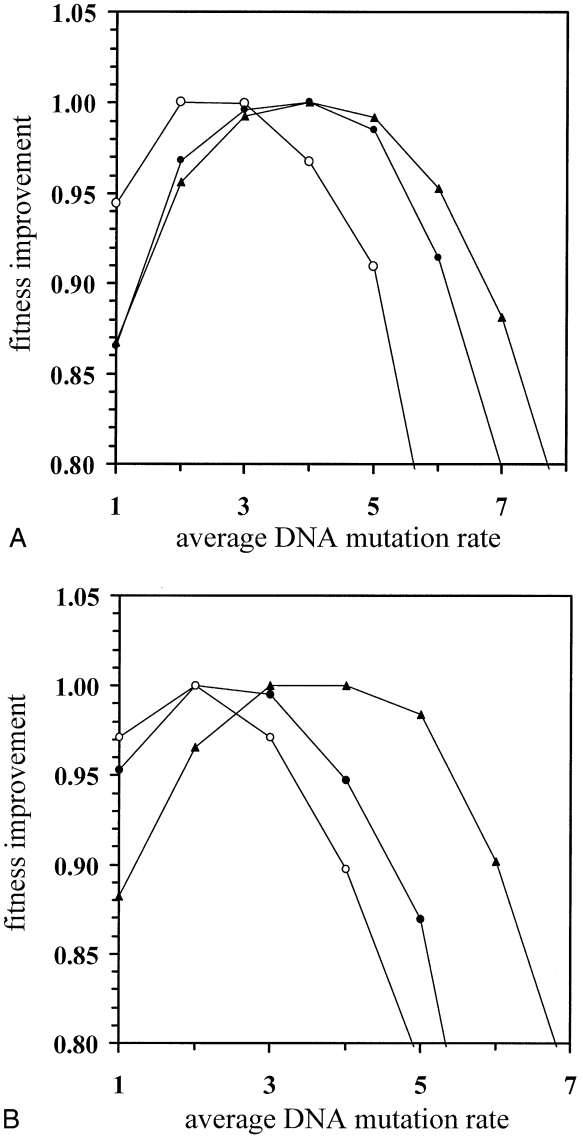


FIG. 6. The optimal DNA mutation rate as determined from a model that incorporates one-body and two-body fitness contributions (similar to a spin glass). The genetic code is included in the model. The data are for a $N = 50$ protein. The fitness improvement is the maximum change in fitness averaged over 10,000 landscapes. To compare the relative location of the optima, the curves have been scaled such that the optima are at 1.0. (a) The optimum mutation rate for the uncoupled landscape as the number of mutants screened increases $M = 1000$ (○), 10,000 (●), and 50,000 (▲). (b) The optimal mutation rate as the landscape ruggedness increases. The number of coupling interactions is 75 (○), 25 (●), and 0 (▲). As the landscape ruggedness increases, the optimal mutation rate decreases. Reprinted from Voigt *et al.* (2000a), with permission.

dues involved in a single coupled interaction, thus assuring that the sequence will not become trapped in a local optimum. However, the finite number of mutants that can be screened imposes an upper limit on the mutation rate. When the sequence is highly optimized, the probability of improving each interaction is small and it becomes harder to optimize a coupled residue. Indeed, it was found that the probability of picking a mutant that has a mutation at a highly coupled residue decreases significantly as the sequence increases in fitness (Fig. 7). This effect intensifies as the number of interactions that are coupled to the mutated residue increases.

4. Dynamic Mutation Rates

It is well known in computer science that steepest ascent is a poor optimization technique for landscapes with many optima. To improve the search, simulated annealing includes an effective temperature that determines the ability of a walker to overcome energy barriers (Kirkpat-

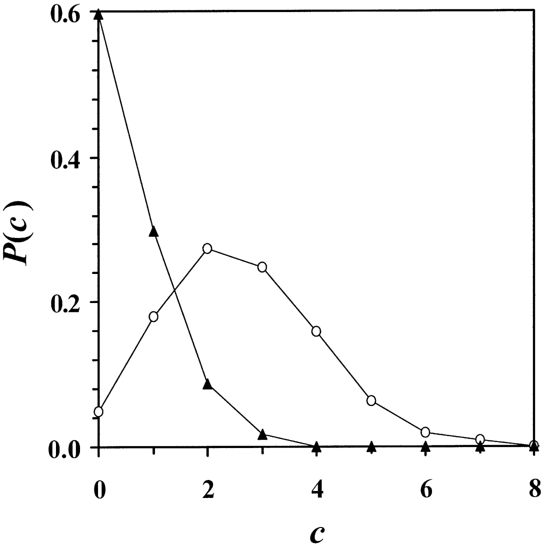


FIG. 7. The probability distribution $P(c)$ of a positive mutation with c coupled interactions occurs as the sequence ascends the fitness landscape (generated using a spin-glass-like fitness function). The distribution is shown at two positions on the fitness landcape, a random sequence (○) and a highly optimized sequence (▲). As the sequence is optimized, the probability that positive mutations will be made at uncoupled residues increases considerably. The mutation rate is an average of one amino acid (three nucleotides) per sequence and the number of mutants screened is 3000. Reprinted from Voigt *et al.* (2000b), with permission.

rick *et al.,* 1983). The simulation starts at a high temperature, allowing a large volume of the search space to be explored freely. Over time, the temperature is decreased (annealed), driving the walker to the optimum. In an adaptive walk, an analog to the temperature is the mutation rate. Initially, a large mutation rate allows a greater sweep of sequence space. As the fitness of the sequence increases, the optimal mutation rate decreases. For genetic algorithms, it has been demonstrated that the outcome is improved when the mutation rate was decreased exponentially over time (Fogarty, 1989; Bäk, 1992; Bäk, 1993).

Consistent with these studies, Husimi and Aita (1998a) found that an initially high mutation rate followed by a slower mutation rate is the most effective search strategy on the uncoupled fitness landscape. Because the probability of finding improved mutations decreases as the sequence becomes more optimized, adaptation via long-jump mutagenesis is rapid at first, then slows. As the coupling between residues increases, the optimal mutation rate decreases more rapidly when the sequence ascends the fitness landscape (Voigt *et al.,* 2000a). The correlation length is short on rugged landscapes, so when the fittest mountain range is found, this information is quickly lost due to the high mutagenesis rate (Kauffman and Levin, 1987).

Mühlenbein (1992) studied the optimal mutation rate for genetic algorithms. By using a Markov chain analysis, he found that the optimal mutation rate p_m is

$$p_m = 1 - \left(1 - \frac{i}{N}\right)^{1/i}, \tag{19}$$

where N is the length of the sequence and i is the number of residues that are not in the optimal state. For $i \ll N$, this reduces to $p_m = 1/N$, that is, there should be approximately one amino acid substitution per sequence for highly optimized sequences. This analysis implies a decreasing mutation rate as the sequence becomes more optimized.

B. Recombination

As a search technique, using mutation and selection alone has several limitations. Evolution via asexual reproduction tends to build up deleterious mutations, ultimately limiting the potential of the experiment, an effect known as Müller's ratchet (Müller, 1964). This effect is exacerbated by high mutagenesis rates, as slightly deleterious amino acid substitutions can hitchhike with positive mutations. Recombination can act to remove neutral and deleterious mutations while allowing the accumulation of multiple beneficial mutations (Stemmer, 1994a). Using recombination also has a practical advantage. More than one mutant may show improved fitness during a single screening effort. Allowing only the fittest mutant to continue to the next generation can be a wasteful strategy. Combining the good mutations onto one gene by recombination has the potential to save experimental effort (Moore *et al.,* 1997). The benefit of recombination also depends on the structure of the fitness landscape. If significant non-additivity is observed, the benefit of combining positive mutations is less certain due to the possibility of combining individually good mutations that together have a deleterious effect.

In the language of genetic algorithms, recombination can exploit information in the mutant population by combining good mutations, but suffers from a lack of exploration (Holland, 1975). Exploration is lost because an amino acid at a specific sequence position cannot be sampled if it does not exist in any of the parents. However, combining good mutations exploits the fact that this diversity has already survived selection (Bergman and Feldman, 1992). Recombination constructs higher order hyperplanes of sequence space from lower order hyperplanes that have higher observed average fitness (Holland, 1975; Spears, 1993). These hyperplanes are referred to as schema (singular: schemata) and represent sets of residues that are collectively beneficial (Fig. 8). During each generation, recombination acts to combine these clusters of good amino acids from different parents onto a single offspring, progressively reducing the offspring to regions of higher fitness (Forrest and Mitchell, 1993). This implies a contradiction for recombination. On one hand, it is better to build up small blocks of fit schema to produce larger schema and ultimately the optimal sequence. However, at some point, the schema become too large and crossover is disruptive (Vose and Liepens, 1991; Eshelman and Schaffer, 1995). For recombination to be successful, the disadvantage of separating beneficial interactions between amino acids has to be less than the advantage of combining good mutations (Otto *et al.,* 1994).

Before we apply the results from genetic algorithm studies to experimental recombination, it is important to understand the fundamental forms of recombination used in each application. There are several recombination methods that have been widely studied for genetic algorithms. Single-point crossover cuts the parents at one location and then swaps the genetic material (Fig. 9a). This is a special case of *n*-point crossover, where *n* is the number of cut points in the sequence. As *n* increases, the fragments become smaller. Diagonal crossover modifies *n*-point crossover to suit multiple parents. For diagonal crossover, each of the parents is cut and reannealed to produce the offspring so that the

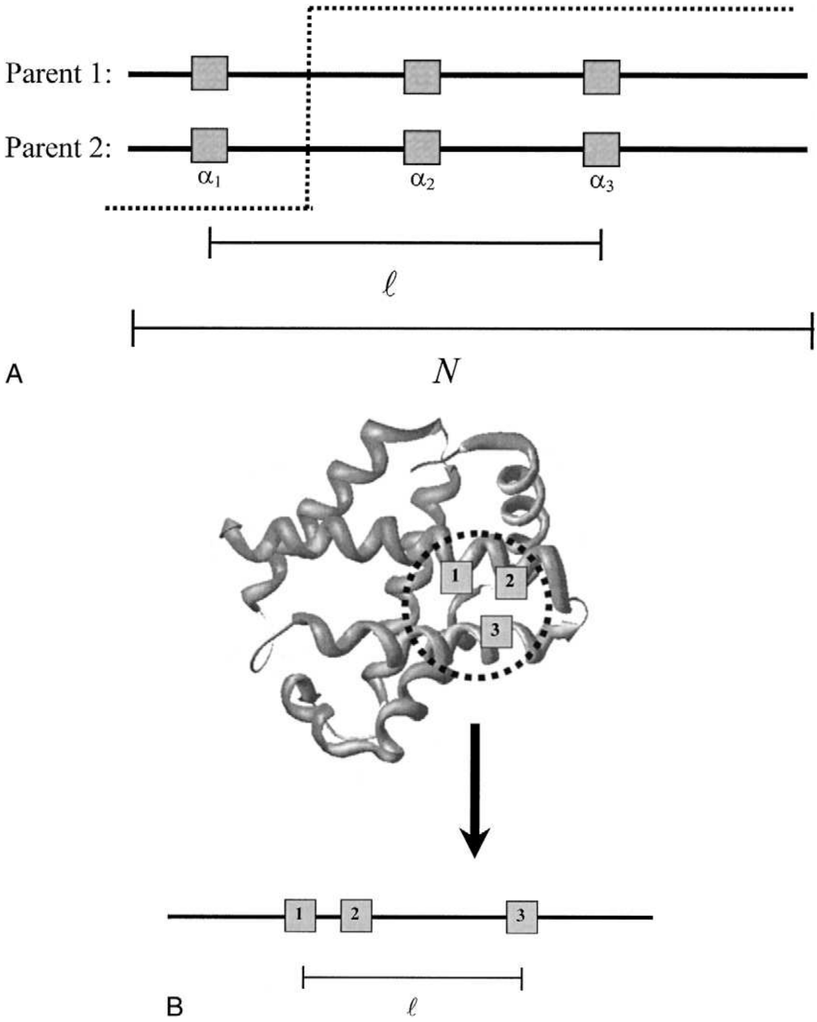


FIG. 8. (a) This figure displays the characteristic parameters in combining schema. The length of the sequence is N and the schemata is made up of the residues α_1, α_2, and α_3. The number of defining bits is three (the schemata is a third-order hyperplane). The defining length is the Hamming distance between the two furthest residues in the schemata and is given as $d(\alpha_1 - \alpha_3) = l$. The dotted line demonstrates a crossover point that disrupts this schemata. Reprinted from De Jong and Spears (1990) with permission, © 1990 by Springer-Verlag. (b) An example of long-range interactions in a real protein leading to a large defining length.

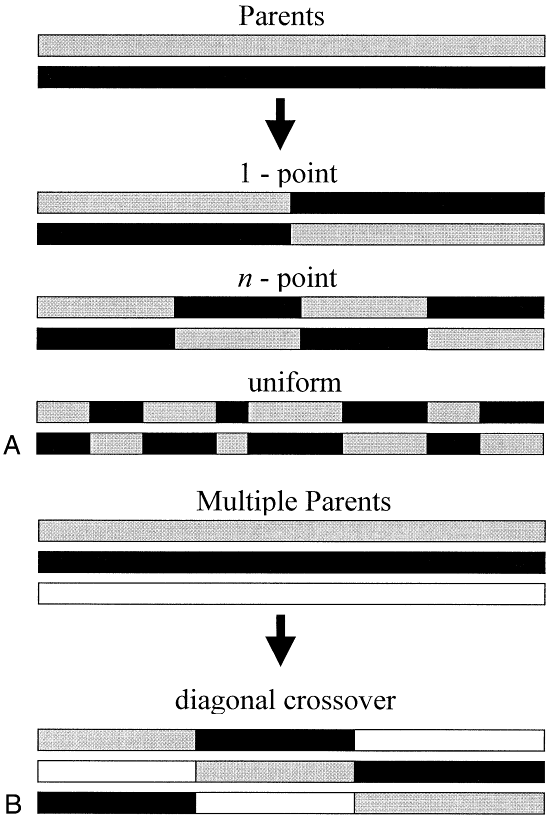


FIG. 9. The types of recombination most commonly utilized in genetic algorithms for (a) one parent and (b) multiple parents. In the cases of 1-point and *n*-point crossover, the cut points do not have to be evenly distributed over the sequence. Uniform crossover gives the offspring the amino acid of either parent with equal probability and can be extended to multiple parents.

parental genes fall along the diagonal [Fig. 9(b)] (Eiben and Schippers, 1996). As a more disruptive technique, uniform crossover allows the offspring to inherit the genetics of either parent at each residue, with equal probability.

In vitro recombination methods include DNA shuffling, random-priming recombination, and the staggered extension process (StEP) (Arnold and Wintrode, 1999). In DNA shuffling, the parental DNA is enzymatically digested into small fragments. The fragments can be reassembled into offspring genes (Stemmer, 1994a; Stemmer, 1994b; Zhao and Arnold, 1997b). In the second recombination method, template DNA sequences are primed with random-sequence primers and then extended by DNA polymerase to create fragments of random length. The template is removed and the fragments are recombined, as in DNA shuffling (Shao *et al.,* 1998). In each of these methods, the number of cut points can be increased by starting with smaller fragments or by limiting the extension reaction. StEP recombination differs from the first two methods because it does not use gene fragments (Zhao *et al.,* 1998). The template genes are primed and extended before denaturation and reannealing. As the fragments grow, they reanneal to new templates and thus combine information from multiple parents. This process is cycled hundreds of times until a full-length offspring gene is formed. Sun (1999) has developed a mathematical model of the distribution of regions that can be reassembled by DNA shuffling and has proposed an application of the theoretical results to the optimal design of experiments.

The parameters to be optimized for experimental recombination include the crossover frequency, the number of parents, and the sequence similarity between parents. Additionally, it is important to understand the conditions under which recombination is useful. For all these questions, the optimal parameters represent a balance between the exploration and exploitation capabilities of the search algorithm. Any process that creates more diversity, such as using many parents, very disparate parents, or small fragment sizes, will improve exploration at the cost of exploitation. The balance between these effects will shift according to the landscape ruggedness and the sampling ability of the mutant library.

1. *Recombination with Limited Sampling*

Because recombination results in a distribution of amino acid contributions from each parent, a minimum screening effort is required to sample all possible combinations. Sampling the offspring that has all the mutations (or the optimal combination) from both parents becomes more difficult as the number of differences increases. Combining many mutations represents the desire for exploration. However, when the library cannot sample all the combinations, the information available from each parent cannot be fully exploited. Thus, recombination should only create the degree of diversity that can be sampled by the mutant library. Similarly, for genetic algorithms, it has been suggested that the best way to size the population is to make it large enough so that it can sample all of the amino acids at each position by recombination alone (Reeves, 1993). The optimal population size would then be able to sample all of the schema shared by the parents, making selection more effective.

An estimation for the required number of screened mutants can be obtained by assuming that the mutations are additive and recombined independently (Moore *et al.,* 1997). The probability that an offspring has d mutations is

$$P_d = \frac{T!}{(T-d)!d!}\left(\frac{1}{M}\right)^d\left(\frac{M-1}{M}\right)^{T-d}, \tag{20}$$

where M is the number of sequences and T is the total number of mutations. When two mutations are combined, only 25% of the library produced by recombination has novel combinations of mutants; the remaining 75% is a mixture of 0-mutant and 1-mutant genes that have been previously observed. The probability of making the rare sequence containing all mutations is only $1/M^T$. To discover the fittest combination of mutations, it is necessary to screen enough recombined sequences to sample the rarest all-mutant sequence; in practice, some oversampling is required. The optimal sequence is not necessarily the rarest, as some of the combinations of mutations may be deleterious through non-additive effects.

2. *The Influence of the Fitness Landscape on Recombination*

The balance between exploration and exploitation shifts toward exploitation as the fitness landscape becomes more rugged. Rugged fitness landscapes have a short correlation length and therefore the accumulation of explorative effects is more rapid. However, there is an upper limit, as extremely rugged landscapes do not allow exploitation (Otto *et al.,* 1994). In the random energy model, the combined effect of two mutations is unpredictable, thus rendering recombination inefficient. The normally exploitative recombination reverts to pure exploration. In this section, we discuss the behavior of recombination on different fitness landscapes with the intention of determining the conditions under which recombination is useful and how the landscape ruggedness affects the experimental parameters.

Exploitation cannot occur unless there are schema that can be utilized by recombination (Kauffman, 1993). In the extreme cases of the uncoupled (smooth) and very rough landscapes, schema are not well defined. Insight into the behavior of schema for different degrees of landscape ruggedness is given by the block model described by Macken and Perelson (1995). They define regions of the sequence as blocks, where each block represents a group of residues that are collectively influenced by selection. If a sequence contains only one block, the model reduces to the random energy model; if the number of blocks equals the number of residues, the landscape is Fujiyama-like (uncoupled). If the blocks are all the same size, then the correlation length of the landscape is equal to the total number of blocks, implying that the uncoupled landscape contains many schema (each residue is its own schemata) and the random landscape contains no schema. Because recombination requires schema to be exploitative, it is expected to fail for either the completely smooth or completely rugged landscapes.

Recombination is a good search strategy when the landscape is of intermediate ruggedness. Recombination is useful when peaks are close in space and thereby contain mutual information of their location in sequence space (Kauffman, 1993; Hordijk and Manderick, 1995). Bornholdt (1998) developed a statistical mechanics approach to solving genetic algorithm dynamics by defining a parent–offspring fitness correlation function (similar to correlation function described in Section II.C.2). He applied this formalism to population dynamics on *NK* landscapes and found that genetic algorithms work most efficiently when the fitness of the parent and the offspring are correlated, as is the case for weakly coupled landscapes. Bergman and Feldman (1992) found low recombination rates are beneficial on rugged landscapes, whereas higher recombination rates are beneficial on smoother landscapes. In summary, recombination is more useful as the landscape is smoother, but is not useful in the limit of the perfectly smooth landscape.

In addition to the landscape structure, the positions of the parents on the landscape are important in understanding when recombination is useful. Stemmer (1994a) postulated that the advantage of recombination over mutagenesis methods is likely to increase as the sequence becomes more optimized. This is because the decrease in the probability of finding positive single mutants makes it harder to sample good mutations. Since identifying a single positive mutant is harder, it is beneficial to combine (exploit) the individual improvements in fitness.

In studies on genetic algorithms, it has been shown that the optimal recombination and mutation rates are not independent; rather, the success of the search algorithm is dependent on their ratio (Pàl, 1995). The effect of mutation is pure exploration. Thus, when the optimal balance between exploration and exploitation is calculated, a large mutation rate indicates that the recombination should be more exploitative. It has been suggested that the mutation rate should be inversely adjusted with the recombination rate (Hesser and Männer, 1990). To study this effect, Kauffman and Macready (1995) calculated the effectiveness of recombination on pooling strategies using the *NK*-model. Cycles of recombination and mutation were applied to small peptides ($N = 6$) with different degrees of coupling between the residues ($K = 0 - 6$). Smooth landscapes required a stronger reliance on mutation and rough landscapes required more recombination. This result is consistent with the notion that more exploration is allowed on smoother landscapes.

3. *Optimal Cutting Strategy*

Recombination has the potential to instigate large jumps in sequence space. However, in doing so, it causes good amino acid combinations to be separated, referred to as crossover disruption (Vose and Liepens, 1991). If two residues are involved in a positive interaction and the cut point for recombination divides the residues onto different fragments, then recombination is disrupting the beneficial interaction. An important concept in modeling disruption is the defining length ℓ of a schemata, which is the distance in sequence between the furthest points in a schemata [Fig. 8(a)]. The number of defining bits is the number of residues involved in the schemata. As ℓ increases and the number of defining bits remains constant, it becomes more likely that a cut will disrupt the schemata. To understand how crossover disruption is realized in proteins, it is necessary to visualize how the schema are distributed (Eshelman and Schaffer, 1995). In protein recombination, the cut points are made on the sequence level, but the schema are on the fitness, or structural level [Fig. 8(b)]. Because the residues that participate in a single schemata are distributed widely in the protein sequence, each schemata is composed of a few defining bits, separated by a large defining length. Two issues are important in optimizing the crossover points: where and how many times the sequence is cut.

If the schema overlap on the primary sequence, they become very difficult to combine. Bogarad and Deem (1999) recognized the need for well-defined cut points in molecular evolution. They modeled directed evolution using a variation on the *NK*-model that includes a notion of secondary structure, referred to as the block *NK*-model. Instead of a single sequence with K interactions between residues, the block model confines the interactions to within segments (the so-called secondary structure) and there are D further interactions across segments. Because the majority of interactions between residues are within a segment, they propose that the natural cut points for recombination are the borders separating segments. Thus, the least amount of crossover disruption is realized. The crossover disruption problems were solved inherently by defining schema in the model. Under these conditions, recombination provided an improvement in the optimization algorithm similar to the combination of good schema in genetic algorithms. However, in real protein systems, the optimal cut points do not necessarily correspond to secondary structure elements and are more subtle and difficult to identify rationally. It has been suggested in genetic algorithm studies that keeping record of the successful cut points from provious generations and preferentially targeting them in future generations aids the search (Schaffer and Morishima, 1987). This approach could be experimentally incorporated by using the restriction enzymes that were previously successful in producing good recombinant mutants. In this way, the combinatorial diversity of the recombinant library could be reduced after each generation.

To determine the optimal number of cut points, it is necessary to return to the balance of exploration and exploitation. Uniform crossover is more exploratory than single-point crossover because the number of schema that can be sampled increases (Spears, 1993). However, the exploration comes at the cost of exploitation because more of the schema are disrupted (Eschelman *et al.,* 1989). The specific balance between these effects will depend on the sampling ability of the library, the structure of the schema, and the ruggedness of the fitness landscape.

The fitness landscape determines the optimal balance between exploration and exploitation and the structure of the schema. Manderick and co-workers ran simulations on *NK* landscapes using the uniform and single-point crossover operators (Manderick *et al.,* 1991; Hordijk and Manderick, 1995). Uniform crossover was the most successful on smooth landscapes, but single-point crossover was more successful on rough landscapes. Rough landscapes have more coupling interactions. A recombination strategy that disrupts the least number of these interactions is more likely to be beneficial. The structure of schema is related to the distribution of coupled interactions between residues. Nearest-neighbor interactions favor the single-point crossover, while random interactions favor uniform crossover (Hordijk and Manderick, 1995). Similarly, when recombining multiple parents, diagonal crossover is better for nearest-neighbor interactions, and uniform crossover is better for random interactions (Eiben and Schippers, 1996).

If long-range interactions are important, the defining length is large; if short-range interactions dominate, the defining length is small. For single-point crossover, a bound can be given for the probability of disruption P_d,

$$P_d \leq \frac{\ell}{N-1}, \tag{21}$$

where ℓ is the defining length and N is the length of the sequence (De Jong and Spears, 1990). The simple probability for even crossover is a very conservative lower bound for the crossover disruption as both the parents could share the same amino acid at a given schemata position. Equation (21) implies that single-point crossover is biased against the successful recombination of schema that have large defining lengths (Fig. 10). Syswerda (1989) demonstrated that uniform crossover usually outperforms one- or two-point crossover in genetic algorithms. If ℓ is small, then the schema are compact and are easily recombined. In this case, it is best for the schema to be preserved by recombination which makes only a few cut points. The average survival probability for a schemata decreases more rapidly for uniform crossover as ℓ increases. When ℓ is large, smaller fragments are more successful because there is a higher probability of combining good schema onto a single gene.

Directed evolution often uses recombination under conditions when crossover disruption is advantageous. The benefits of disruption are realized late in evolution, when the population diversity is low, and when the number of parents that can be sampled is small (De Jong and Spears, 1990). Under the influence of strong selection, positive mutants are worth more and therefore exploitation is more beneficial, thus allowing more disruption (Bergman and Feldman, 1992; Prügel-Bennett and

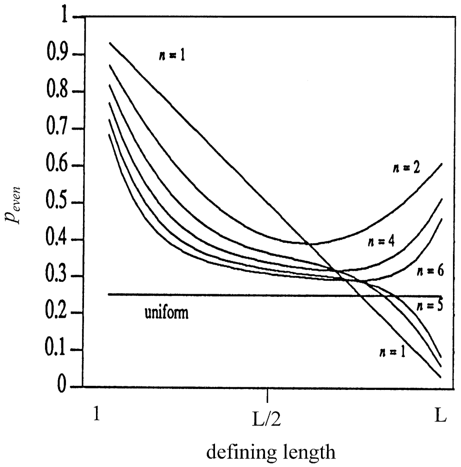


FIG. 10. The probability of disrupting a schema versus the defining length for a sequence of length N. The lines are for n-point crossover and the flat line is for uniform crossover. Note that for large defining lengths, uniform crossover is less likely to be disruptive. Reprinted from De Jong and Spears (1990) with permission, © 1990 by Springer-Verlag.

Shapiro, 1994; Prügel-Bennett and Shapiro, 1997). When the sequence similarity between parents is high, crossover is more disruptive and is more likely to create a diverse library (Spears and De Jong, 1991). Parents that are too close in sequence space tend to produce offspring that preserve suboptimal schema. A higher disruption rate avoids this problem. When the parents are on local optima, it was found that the highest disruption produced the best results (Pàl, 1995). These observations describe a situation that is very similar to that encountered in directed evolution experiments. Taken together, the evidence suggests that crossover with many cut points is useful, although if the landscape is expected to be particularly rugged, fewer cut points may be beneficial.

4. Number of Parents and Their Sequence Similarity

The recombination of genes that differ only by a few point mutations is primarily a technique to combine good mutants and save experimental effort. Although recombination reduces the number of mutants that need to be screened, most of these moves would be attainable through point mutagenesis alone in future generations. Recombination is a powerful technique to search large areas of sequence space that cannot be sampled with point mutagenesis in a reasonable number of generations. The difficulty with this procedure is similar to the problem with high mutagenesis rates: Most of sequence space is empty of function. To avoid this problem, the initial parents have to contain beneficial schema that can be combined onto a single sequence.

Stemmer and coworkers have demonstrated that the recombination of schema can be achieved experimentally by picking initial sequences that have evolved independently in different species (Crameri *et al.*, 1998; see chapter by Ness *et al.* in this volume). They shuffled four genes that had 59% to 82% amino acid sequence similarity and found that the fitness improvement was significantly greater than that attained by point mutagenesis and local recombination. Although this ''family shuffling'' searches a large volume of sequence space, the sampling within this volume is sparse. Additional rounds of point mutagenesis can then be used to find the local optimum. Parameters that arise out of this experimental technique that depend on the structure of the fitness landscape include the optimal number of parents and their degree of sequence identity.

These questions have been investigated by Eiben and coworkers and Coker and Winter, who found that increasing the number of parents increases the total fitness improvement and increases the speed of the optimization (Eiben *et al.*, 1995; Coker and Winter, 1997). However, by increasing the number of parents, the disruption caused by recombination also increases. Recombining many parents has an effect similar to recombining two parents that have many mutations. If it is possible to sample all the possible mutation combinations, recombining all available parents is advantageous. However, this is often experimentally infeasible, so a compromise between the difficulty of screening and the diversity of more parents has to be reached. To illustrate this, Moore *et al.* (1997) assumed that the mutations are beneficial and can be combined independently and additively and simulated the effects of recombining multiple parents when screening is limited. By simulating the recombination of ten possible parental genes, they demonstrated that, with finite sampling, higher fitnesses could be achieved with five parents than with all ten.

By using the *NK*-model, Eiben and Schippers (1996) studied the effect of allowing many parents in the genetic algorithm as a function of the landscape ruggedness. The best performance was always achieved using multiple parents. However, the benefit of adding parents died off quickly after six to eight parents for intermediate landscape ruggedness. As the landscape becomes more rugged, the relative gain from multiple parents decreases, as distant parents share less fitness information.

Similar to the number of parents, the sequence similarity between parents requires a balance between exploration and exploitation. If the sequence similarity is high, mutations can be readily exploited, but at the cost of exploration. As expected, this balance will shift toward exploration when the number of mutants that can be screened is limited. More interestingly, landscape theory predicts the optimal similarity based on the fitness correlation between parents. The correlation length decreases as the landscape becomes more rugged; therefore, the parents have to be closer in sequence space (Manderick *et al.*, 1991). Sumida *et al.* (1990) studied the effect of the landscape ruggedness on the optimal genetic diversity between the parents. A genetic algorithm was modified so that the population was divided into small, isolated subpopulations. For landscapes that are extremely rugged or smooth, the benefit of subdividing the population was small. For landscapes of intermediate ruggedness, it was advantageous to maximize the subdivision.

There have been several studies with genetic algorithms to probe the effects of forcing diversity among the parents prior to recombination. Evolutionary search can be improved by preventing recombination between sequences that share high sequence identity (Eshelman and Schaffer, 1991; Pàl, 1995). Only parents separated by a Hamming distance above some threshold were recombined, and improvement was demonstrated on a variety of test functions and the spin-glass landscape. Rather than using the Hamming distance metric, Craighurst and Martin (1995) defined the genetic distance through the ancestry of the parents.

They found that prohibiting crossover between parents that are too close in ancestry improves the outcome of the search. The benefit of preventing phylogenetic incest is high for low mutation rates and then, at a critical mutation rate, reverses so incest prevention is deleterious. Because mutation is explorative, a low mutation rate was required to balance the extra exploration caused by incest prevention.

C. *Quantifying Diversity*

Experimental limitations impose significant restrictions on the number of mutants that can be screened. Inventing a screen that both captures the essence of a complex chemical function and can be repeated for thousands of mutants is a difficult first step for *in vitro* evolution (Zhao and Arnold, 1997a). Because much of the experimental effort is devoted to screening, there is motivation to ensure that the cost of screening is minimized. Because of practical size limitations, the content of the screening library needs to be optimized; if the library cannot be bigger, then it should be smarter. For instance, biased mutational methods can produce a library that is more likely to contain stable mutants. As another strategy, some methods are designed to quickly accumulate positive mutations, at the risk of ultimately converging to a suboptimal solution. In this section, we explore advantages and disadvantages of these approaches.

1. *Optimal Screening Library Size*

It is possible to quantify the fitnesses of a million mutants for a single generation (Daugherty *et al.,* 1998; Joo *et al.,* 1999). This is impressive as this library size can easily sample the roughly 5600 possible single mutants of a 300-residue protein. However, it samples only a fraction of the 1.6 million possible double mutants and far less than the 3 trillion triple mutants. When is screening a single library of 10^6 better than 10 generations of 10^5 mutants? Although the improvement in fitness increases with the size of the screening library, the benefit of accumulating positive mutations over multiple generations is compromised. Both experimental and theoretical studies have suggested that the best method may be short, adaptive walks utilizing small libraries (Chen and Arnold, 1993; Matsuura *et al.,* 1998).

A theoretical approach to determine the required number of screened mutants is based on the landscape paradigm. The number of uphill paths on the landscape given the fitness of the sequence is related to the library size required to probabilistically capture a single path. Using the uncoupled fitness function, Aita and Husimi (1996) analytically determined the distribution of single-mutation fitnesses around a walker as it ascends to a peak. For a sequence of N residues that is distance d_p from the global optimum, the average fraction of fitter mutants $<\Theta_p>$ at Hamming distance d was calculated (Fig. 11). When $d = 1$, the average mutant distribution can be estimated as $<\Theta_p(d = 1)> \approx d_p/$

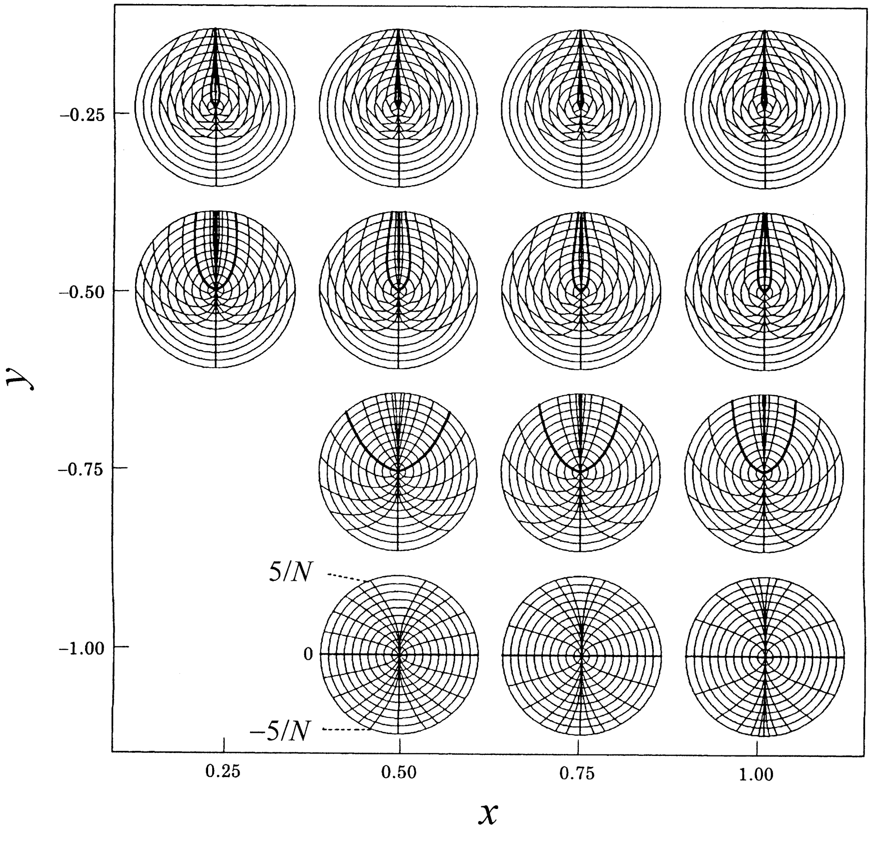


FIG. 11. The local fitness distributions around fourteen representative wild types. The curves were determined analytically for the fully additive landscape by Aita and Husimi for sequence length $N = 60$ and alphabet size $A = 20$. Each wild type is shown at the center of the concentric circles. The axes y is the scaled fitness ($\equiv F/|\varepsilon N|$, ε is the mean of F and here is negative) and x is the scaled Hamming distance from the optimum ($\equiv d_0/N$). Each local fitness distribution is expressed as a concentric pie chart showing the fraction of mutants having Δy between l/N and $(l + 1)/N$, where $l = -5, -4, -3, \ldots, 4$. The thick curves represent the contours satisfying $\Delta y = 0$. Reprinted from Aita and Husimi (1998a) with permission, © 1998 by Academic Press.

($2N$), predicting that the number of mutants that need to be screened in order to find fitness improvements increases linearly as the sequence moves closer to the global optimum.

In the *NK*-model, as K increases, the number of fitter neighbors decreases more quickly as the sequence becomes more optimized (Kauffman and Weinberger, 1989). Thus, in order to discover improved mutants, the number of mutants screened has to increase more rapidly on random landscapes as the sequence increases in fitness (Fig. 4). The rate of decrease for the number of uphill paths is greater for rugged landscapes due to the shortening of the walk length to local optima. This implies that a protein that is tolerant (a smoother landscape) can undergo more rounds of mutation and improvement.

Macken *et al.* (1991) identified two behaviors of an adaptive walk as it ascends the fitness landscape that are separated by a distinct phase transition. As a sequence becomes more optimized, the number of fitter mutants decreases. At some point, it becomes difficult to sample a fitter mutant given the number that can be screened. The probability of a sequence with fitness F having a single-mutant neighbor with lower fitness is $G(F)$. The probability that a sequence with D single-mutant neighbors is at a local optimum G^{D-1} (F) is approximately

$$\begin{aligned} G^{D}(F) &\approx \exp[\ln(G^{D})] \\ &= \exp[-D(1-G) - D(1-G)^{2}/2 -. .], \end{aligned} \tag{22}$$

which defines a region $D(1 - G) \gg 1$, where G^D is negligible. Walks that start outside of this region obtain large fitness improvements in a few mutational steps. After the dramatic initial gain in fitness, the subsequent mutational steps tune the fitness near the optimum. A phase transition occurs at $D(1 - G) = 1$, where the behavior of the walk changes from that of the outside region to that of the inside region. Typically, in directed evolution experiments, $G \approx 0.01 - 0.001$ and $D \approx 1800 - 3000$ (taking into account the genetic code), so $D(1 - G) \approx 1.8 - 30$, indicating the directed evolution starts in the outer region. When the inner region is reached, the search becomes exponentially more difficult for each additional uphill step.

Macken and Perelson studied antibody affinity maturation as a random walk on the random energy landscape (Macken and Perelson, 1989; Macken *et al.,* 1991; Macken and Perelson, 1991). The total number of mutants tried before a positive mutation is discovered $T(F)$ is a measure of the change in the necessary size of the mutant library. The expected value of T, given that F is not a local optimum, is derived as

$$E[T(F)] \approx D\frac{e^{-\theta}}{1-e^{-\theta}}\int_0^{\theta}(e^z-1)\frac{dz}{z}. \tag{23}$$

For the inner region $\theta \ll 1$, Eq. (23) reduces to $E[T(F)] \approx 1/(1 - G(F))$. The total number of mutants tried increases exponentially as the sequence approaches the global optimum (Fig. 12). Outside of the boundary, the mean number of trials approaches the asymptote of $\approx 0.78D$, indicating that the required mutant library size is on the order of the dimensionality of the landscape (Macken *et al.,* 1991).

Van Nimwegen and Crutchfield (1999a) studied the optimum population size to minimize the required number of fitness evaluations E before

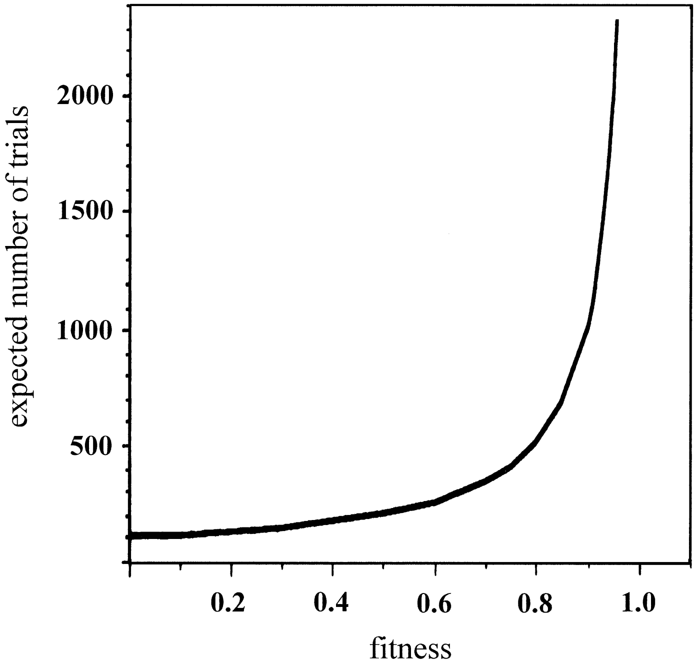


FIG. 12. The average number of mutations tried at a suboptimal fitness plotted versus the fitness for the case $D = AN = 1500$, where A is the alphabet size and N is the sequence length. Note the rapid increase in the number of required trials when the sequence reaches the inner boundary region. The line is generated analytically under the assumption of the random energy landscape. Reprinted with permission from Macken, C. A., Hagen, P. S., and Perelson, A. S., Evolutionary Walks on Rugged Landscapes, SIAM *J. Appl. Math.,* **51** (1991), p. 821. Copyright © 1991 by the Society for Industrial and Applied Mathematics. All rights reserved.

an optimum is reached. As the population size increases beyond a threshold, E increases very slowly, implying that it is important that the population stays above the minimum size, but less important to stay below the maximum size (Fig. 5). Because the maintenance of a large population is difficult in directed evolution experiments, these results indicate that the optimal number of mutants screened is as close to the lower bound as possible. This point is just high enough to avoid the problems of having too small a population.

As the noise in the fitness measurement increases, more mutants need to be screened to discover positive mutants. By modeling these effects, Rattray and Shapiro (1997) calculated the optimal population size to achieve the greatest fitness improvement for a genetic algorithm. If M_0 is the population size for zero noise, β is the selection strength, and σ is the standard deviation of the noise, then the new population can be scaled as

$$\frac{M}{M_0} \sim \exp[(\beta\sigma)^2] \tag{24}$$

to remove the effects of noise. This population resizing is valid as long as the selection strength and noise are uncoupled. Equation (24) implies that very small increases in the noise will dramatically increase the required population size.

2. *Statistical Mechanics and Information Theory*

The optimal size of the mutant library is a parameter that has immediate physical meaning. A more challenging parameter to define is the quality of the library. For example, quality can be broadly interpreted as the structural robustness or the fraction of positive mutants. The short-term goal is to produce a library that maximizes the probability of finding fitter mutants and the long-term goal is to maximize the total fitness improvement after multiple generations. Research in statistical mechanics and information theory has introduced methods that can be used to quantify the quality of a mutant library.

The observation that some sequence positions are more tolerant to mutation initiated the application of information theory as a method to understand the importance of these residues to the structure and function of the protein (Reidhaar-Olson and Sauer, 1988). A residue that is intolerant to mutations has high information content, whereas a tolerant residue has low information content. By fixing the amino acid identities of residues as the fitness is improved, optimization acts to increase the information content of the sequence (Dewey and Donne, 1998). The information content of the mutant library can be used as a measure of a method's ability to generate diversity (Strait and Dewey, 1996).

The Shannon information content (in thermodynamics, the entropy) can be calculated from the probability distribution of allowed amino acids substitutions (Fontana and Shuster, 1987; Saven and Wolynes, 1997; Dewey and Donne, 1998). Counting the number of sequences at a given fitness is mathematically isomorphic to calculating the entropy at a given fitness, $S(F) = k_B$ In Ω, where the number of states Ω is the number of sequences at fitness F. The entropy s_i for a given site i can also be calculated from Eq. (25)

$$s_i = - \sum_{\alpha=1}^{m_i} p_i(\alpha) \ln p_i(\alpha), \tag{25}$$

where m_i is the number of amino acids allowed at site i and $p_i(\alpha)$ is the probability that α exists at i. A small value of s_i represents the conservation of identity and a large value of s_i indicates high mutability. This local sequence entropy captures the structural constraints on the amino-acid identity at certain sites (Saven and Wolynes, 1997).

To specifically model protein tolerance, Aita and Husimi (1996) introduced an entropy-based method, using the fitness change w that is induced by each amino-acid substitution. The advantage of this formulation is that the fitness difference is an experimentally observable quantity, whereas the probabilities required for Eq. (25) are not. The information entropy I_j of site j is

$$I_j = \log_2 \frac{A}{Z_j \exp[-\langle w_j(\alpha_{\mathrm{j}})\rangle]} \tag{26}$$

where Z_j is the partition function,

$$Z_j \equiv \sum_{\alpha} \exp[w_j\,(\alpha)] \tag{27}$$

and

$$\langle w_j(\alpha)\rangle_j \equiv \frac{1}{Z_j} \sum_{\alpha} w_j\,(\alpha) \exp[w_j\,(\alpha)]. \tag{28}$$

If I is large, then the amino acid found at that position is highly conserved. Conversely, a low I indicates that many amino acids are allowed at that site. This information content is directly related to the Shannon entropy through a Boltzman weighting of the fitness changes $w_i(\alpha)$ to calculate the probability that a residue is in amino acid state $p_i(\alpha)$.

One way to improve the information content of the mutant library is to only mutate sites that are predicted not to severely disrupt the function or stability (high entropy sites). For example, based on the requirement for a protein to fold, the mutant library could be improved by constraining mutations to conserve hydrophobicity. Following this idea, Hecht and co-workers combinatorially designed an α-helical protein by constraining the core and surface regions to be hydrophobic and polar, respectively (Kamtekar *et al.,* 1993). This created a biased library: Some design information was included so that a larger fraction of the mutant sequences folded properly. It should be noted that conservation of polarity is not always observed (Hellinga *et al.,* 1992) and the biased library produced by Hecht and co-workers is neither functional nor optimized. In directed evolution studies on pNB esterase, no discernable rules for substitution emerged, including hydrophobicity, size, and charge (Spiller *et al.,* 1999).

From the observation that the directed evolution algorithm tends to find positive mutations at relatively uncoupled positions (Section III.A.3), it follows that the uncoupled regions of the protein should be targeted for mutation (Voigt *et al.,* 2000b). Because the number of mutants that can be screened is limited, there is an upper bound on the degree of coupling where positive mutations will occur. More highly coupled residues require a rearrangement of amino acids that is unlikely given the limited mutation rate. It is extremely unlikely that a mutation will be observed at a residue with more coupled interactions than the upper limit of this range. Avoiding the regions of high coupling decreases the total number of residues undergoing mutagenesis. The same library size is then more effective at searching the reduced space, thus increasing the upper limit of the degree of coupling. Not allowing mutagenesis at the most highly coupled residues actually increases degree of coupling that can be sampled by directed evolution.

We chose subtilisin E to test our prediction that directed evolution makes mutations at uncoupled positions (Voigt *et al.,* 2000b). Directed evolution increased the temperature optimum for activity, T_{opt}, of *Bacillus subtilis* subtilisin E from 59° to 76°C, with eight mutations (Zhao and Arnold, 1999). In an independent study, thirteen mutations improved the activity toward the hydrolysis of succinyl-Ala-Ala-Pro-Phe-*p*-nitroanilide (s-AAPF-*p*Na) in the organic solvent dimethylformamide (DMF). The mutants were found by screening 2000 to 5000 clones from a mutant library generated with a average mutation rate of two to three nucleotide substitutions in each generation. This mutation rate produces about one amino acid substitution per sequence.

Based on a rotamer and force field description of the interactions between residues, mean-field theory can be used to estimate the probabilities required by Eq. 25 (Ponder and Richards, 1987; Dunbrack and Karplus, 1994; Lee, 1994; Koehl and Delarue, 1996; Dahiyat and Mayo, 1997; Saven and Wolynes, 1997; Voigt *et al.,* 2000b). A tabulation of the entropy at each position produces an entropy distribution (Fig. 13, see Color Insert). The adaptive mutations are strongly biased toward highly tolerant positions, suggesting that retaining structural stability is important in evolving other enzyme properties. The information from the structural entropy calculations can be incorporated in several experimental methods (Voigt *et al.,* 2000b). First, site saturation mutagenesis can be applied at positions that are predicted to be the most tolerant. The positive mutants are then recombined using DNA shuffling to compound the fitness improvement. As a second method, a portion of the gene that is determined to have an above average total tolerance can be targeted using cassette mutagenesis. Rather than restricting the search to single sites, this allows a massive combinatorial search of a region that has been predicted to be able to withstand the additional diversity. Finally, the experiment can be based on the concept of family shuffling, where the genes from divergent species are recombined (Crameri *et al.,* 1998). In family shuffling, the seqeces have previously survived natural selection; thus, the high diversity between the sequences is less likely to have a deleterious effect on the structure and function. The entropy profile of a structure predicts the positions that are essential to maintain the structure, allowing the tolerant sites to be mutated *en masse* to produce a chimera family of artificially divergent sequences. Recombining the artificial chimera sequences produces a mutant library with large sets of mutations that have been calculated not to disrupt the structural integrity.

3. Analyzing the Mutant Fitness Distribution

When the mutants are screened, the information is generally only used to determine the top sequences that become the parents to the next generation. During this process, a large number of fitness data are generated. Because sequencing is time consuming and expensive, each fitness cannot be assigned to a sequence. The lack of sequencing data means that only the probability distribution of offspring fitnesses can be analyzed. In current practice, this information is largely ignored; however, it may provide useful insight into the structure of the fitness landscape, thus guiding the setting of experimental parameters. The information in the mutant library can be used to analyze the properties of the fitness landscape in theoretical models and genetic algorithms, leading to predictions of the behavior of a random walk.

Urabe and coworkers developed a model that captures additive and non-additive mutational effects in directed evolution and fit their results to the experimental fitness distribution of catalase I (Matsuura *et al.,* 1998). By investigating the degree of non-additivity of specific properties, they tuned the parameters of the experiment to suit the fitness landscape. They found an average degree of non-additivity that seems to be constant over various functions (catalase activity, peroxidase activity, thermostability). The notion of an average degree of non-additivity may indicate that the fraction of highly coupled residues is roughly constant for globular proteins. Matsuura *et al.* used the average degree of non-additivity as a predictive tool to estimate the mutant fitness distribution of the next round of evolution. They employed this method to develop a proactive evolutionary strategy in which the data produced at each generation are analyzed to predict the most appropriate mutation rate and screening library size for the next round (Fig. 14).

Aita and Husimi (1996) proposed that the additive model can be applied to give a rough estimate for the Hamming distance from the wild type to the optimum, the fitness slope near the wild type, and the height of the optimum. They calculated the expected fitness distribution and compared this to experimental data produced by the mutagenesis of a region of *E. coli* lac promoter. Based on a fit between the theoretical and experimental distributions, they estimated that the Hamming distance between the wild type lac promoter and the optimum is seven to ten amino acid substitutions and the activity could be improved 100 to 1000-fold.

The strategy of recursive ensemble mutagenesis (REM) uses information gleaned from previous rounds of evolution to search sequence space more efficiently (Arkin and Youvan, 1992; Youvan *et al.,* 1992; Delagrave and Youvan, 1993). A small number of initial mutants are screened and analyzed, allowing the next, larger round of mutagenesis to be more productive. First, a group of sites is picked to undergo random mutagenesis. Next, this mutant library is screened and the positive mutants are isolated. If, for example, only hydrophobic residues at a position cause positive mutants, then this information can be used to create the next mutant library in the cycle. As the diversity of possible amino acids decreases at a each site, the entropy of that site decreases. Youvan and coworkers successfully used this method to simultaneously alter five to

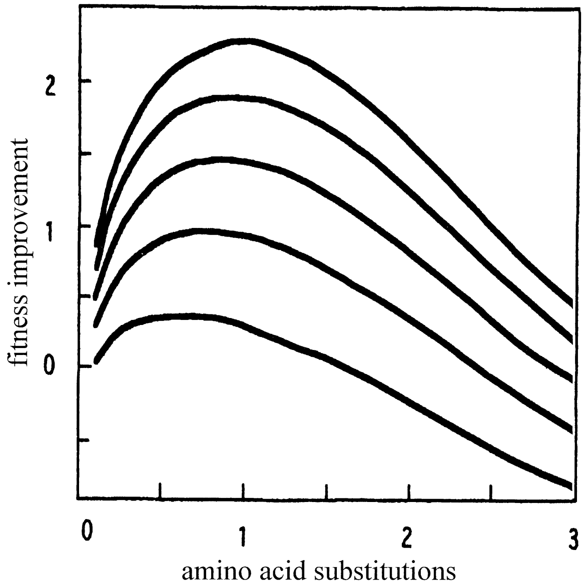


FIG. 14. Predicted correlation between the improvement in fitness achieved *W* and the average mutation rate (Np_m) for different numbers of mutants screened for the D130N mutant of catalase I. The number of mutants screened (from top to bottom) is: 3 × 10^5, 3 × 10^4, 3 × 10^3, 3 × 10^2, and 3 × 10. A clear optimum is observed for all the screening library sizes at around $Np_m = 1$. Reprinted from Matsuura *et al.* (1998) with permission.

nine out of thirteen residues. The primary weakness of the REM method is the extensive sequencing required, which significantly slows the search.

Predictive models based on the mutant fitness distributions have been proposed for genetic algorithms. Grefenstette (1995) found a strong correlation between parent and offspring fitnesses for a library of test landscapes. As the mutation rate was increased, the correlation diminished. The relationship between the mutation rate and the parent-offspring fitness correlation can be used to gain insight into the structure of the fitness landscape. Prügel-Bennett and Shapiro (1994) developed a statistical mechanics approach to understanding genetic algorithms and applied it to evolution on a spin-glass landscape. From this formalism, they were able to determine the optimal degree of selection, based on the fitness correlation between the mutant and parent fitnesses *m*. They obtained relationships, based on *m*, between the mean, the standard deviation, and the higher-order cumulants before selection and the cumulants after selection.

Mean-field theory can be used to predict the effects of mutation rate and parent fitness on the moments of the mutant fitness distribution (Voigt *et al.,* 2000a). In this analysis, only the portion of the mutant distribution that is not dead (zero fitness) or parent (unmutated) is considered. The mutant effects are averaged over the transition probabilities without the cases of mutations to stop codons or when no mutations are made on a sequence. In order to obtain the fitness distribution, two probabilities are required: (1) the probability $p_i(a)$ that a particular amino-acid identity a exists at a residue i, and (2) the transition probability that one amino acid will mutate into another $Q = 1 - (1 - p_m)^3$. The probability vectors $p_i(a)$ can be determined through a mean-field approach (Lee, 1994; Koehl and Delarue, 1996; Saven and Wolynes, 1997). The amino acid transition probabilities Q are calculated based on the special connectivity of the genetic code and the per-nucleotide mutation rate. Removing transitions to stop codons and unmutated sequences only requires the proper normalization of the probabilities p_i and the moments. For example, the first moment of the fitness improvement w of the uncoupled fitness function is written as

$$\langle w \rangle = C\sum_{i=1}^{N}\sum_{a}\sum_{b} [f_i(a) - f_i(b)]P_i, \tag{29}$$

where a is the wild type amino acid identity at residue i and b is the amino acid identity in the mutant sequence. Equation (29) is normalized by

$$C = \frac{1}{1 - \left(\dfrac{1 - Q}{1 - \dfrac{3}{63}Q}\right)^{N}}, \tag{30}$$

as well as the probabilities

$$P_i = \frac{p_i(a)\left[\dfrac{3}{63}Q + (1 - Q)\delta_{a,b}\right]}{\left(1 - \dfrac{3}{63}Q\right)}, \tag{31}$$

such that the unmutated and stop codon sequences are removed from the distribution. The term $\delta_{a,b}$ equals one if $a = b$ and $\delta_{a,b}$ equals zero if $a \neq b$.

The mean-field moments up to the second order (average and standard deviation) were derived for a two-body coupled fitness function. The change in the fitness distribution was captured as the sequence ascends the fitness landscape (Fig. 15). By increasing the number of two-body coupling interactions between residues, the effect of the landscape ruggedness on the moments was calculated. As the fitness landscape becomes more rugged, the second moment increases and the third moment decreases. In other words, as the sequence ascends the fitness landscape, the mutant distribution of a highly coupled sequence spreads out (diffuses) and becomes skewed toward less fit mutants (drifts) more quickly. In addition, the dependence of the moments on mutation rate can be predicted by recalculating the transition probabilities Q. As the mutation rate increases, both the drift and the diffusion of mutants from the parent increase. Because rugged landscapes have less correlation, these effects are exaggerated as the coupling between residues increases. Through these approaches, it may be possible to model the mutant fitness distribution to experimentally obtain statistical parameters that describe the fitness landscape.

4. *Premature Convergence*

Although the evolutionary parameters can be optimized to expedite genetic search, this acceleration can cause premature convergence onto a local optimum (Schaffer *et al.,* 1991; Forrest, 1993). The population needs to retain diversity so that a wide sampling of sequence space can be made. Each time a residue is fixed in the population (a loss of entropy), this represents a limitation in the search. If convergence is too rapid, then many residues become fixed prematurely, leading to a suboptimal solution. There is an optimal rate for the loss of diversity that allows the population to converge quickly while minimally affecting the quality of the final solution (Baker, 1985). This can be achieved by varying the evolutionary parameters with the intention of slowing the convergence, thus slowing the loss of entropy. Many modifications to the genetic algorithm have aimed at solving the problem of premature convergence. On the other hand, premature convergence may be a good quality for an *in vitro* evolution search. If the number of generations is limited, the adaptive walk should be as rapid as possible. The optimum convergence rate balances speed versus the quality of the final solution. Following this strategy, premature convergence has been implemented

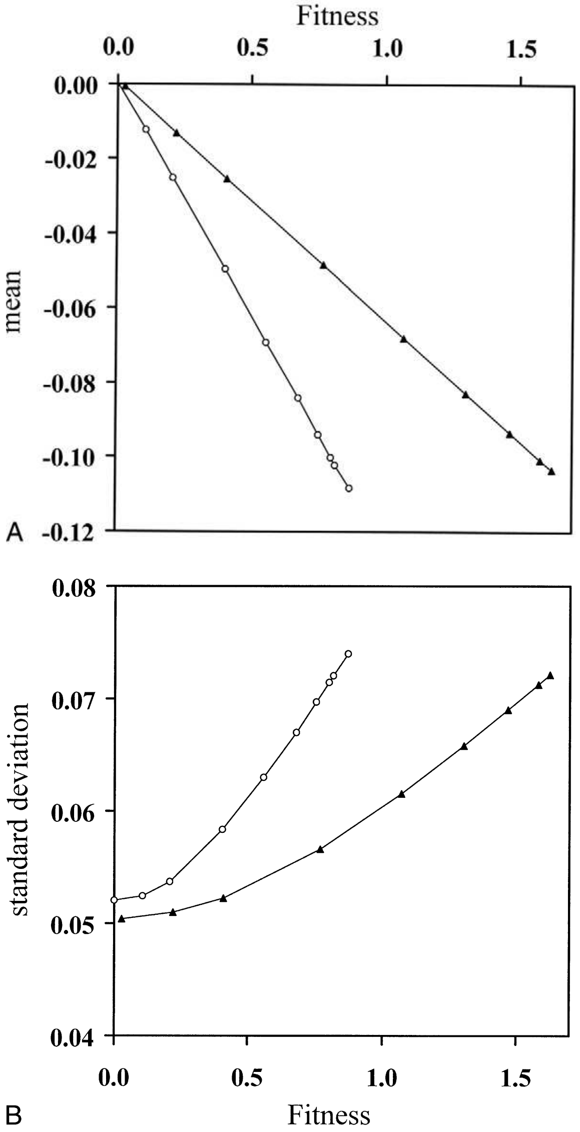


FIG. 15. The (a) mean and (b) standard deviation of the mutant fitness distribution as the sequences ascends the fitness landscape. The data are shown for a smooth (▲) and rugged (○) landscape.

experimentally in the form of pooling algorithms and a phage display model.

In pooling algorithms, pools of sequences are constructed so that each pool shares a specific subset of sequences that does not overlap with the distribution of sequences in other pools (Geysen *et al.,* 1987; Houghten *et al.,* 1991). Fixing the amino acids at specific positions for each pool creates the subset. The pools are then tested for activity and the best pool is picked, which becomes the parent to a new population of pools. Cycles of picking and dividing pools are continued until the best pool contains only one sequence. The advantage of a pooling algorithm is that it allows a greater fraction of sequence space to be explored because many sequences are being screened at once. However, fixing the sequence positions reduces the sequence entropy faster than natural selection alone, causing premature convergence (not necessarily even to a local optimum). Using the *NK*-model, Kauffman and Macready (1995) demonstrated that as the landscape ruggedness increases, the penalty for losing entropy quickly increases, indicating that premature convergence is more deleterious for rugged landscapes. These results suggest that pooling algorithms work best for smooth landscapes (such as thermostability) and are poor for rugged landscapes (such as the antibody V-region).

Phage display is widely used to discover polypeptides that bind strongly to various targets (Irvine *et al.,* 1991; Mandecki *et al.,* 1995; Levitan, 1997). The specifics of modeling phage display are not in the scope of this chapter and have been extensively reviewed elsewhere (Levitan, 1997). Levitan developed a stochastic model, based on kinetics, to optimize the selection parameters of phage display (Levitan, 1998). Levitan studied the effect of increasing the selection strength after each round of washing. The selection strength was increased by decreasing the number of target molecules on the column by a factor of g after each selection round. When $g = 1$, the selection strength is held constant, which insufficiently enhanced the concentration of the high-affinity phage. However, if the rate of increase in the selection strength was too rapid ($g \ll 1$), then the isolation of the high-affinity phage was not pronounced over background. A relatively slowly increasing selection strength required more rounds of selection to obtain the same isolation over background. The setting of the parameter g is an example of balancing premature convergence (Fig. 16). On one hand, the number of generations can be reduced by increasing the rate of selection. However, if the rate is overly rapid, no selection will occur.

Theoretical studies have been widely applied to genetic algorithms to study the control of premature convergence. All experimental parame-

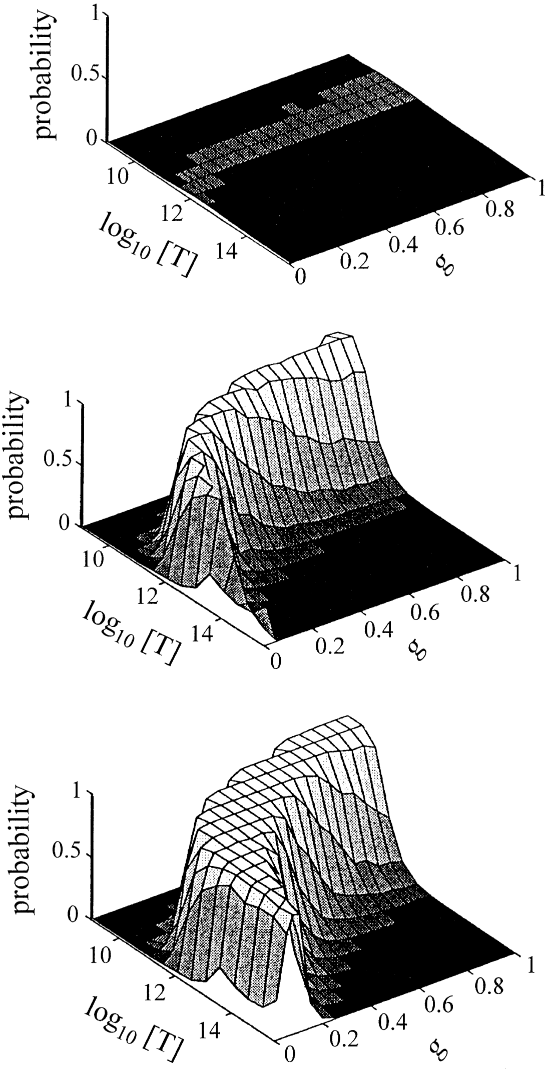


FIG. 16. Probabilistic performance measure for finding the top-ranked phage, where phage are ranked by decreasing target affinity. Individual plots are functions of the initial concentration of target molecules [T^{tot}] and the stringency parameter *g*. Calculations were also performed by Levitan to determine the probability of picking *r* of the top-ranked phage (data not shown). From top to bottom, the plots represent the results after two, four, and seven iterations of selection. Note that after seven iterations, there is a wide range of optimal parameters. Reprinted from Levitan (1998) with permission, © 1998 by Academic Press.

ters can be set with the goal of either accelerating the search or allowing a slow convergence to a better solution. An initially high mutation rate, followed by a slow mutation rate, avoids premature convergence because both local and global mutation moves are allowed (Schober *et al.,* 1993; Aita and Husimi, 1998a). Recombination also contributes to premature convergence. Single-point crossover increases the number of deleterious hitchhiking mutations while uniform crossover disrupts good schema. If recombination is not adequately stringent to remove hitchhikers, then the population will prematurely converge onto a suboptimal solution (Schaffer *et al.,* 1991). Premature convergence can be avoided by forcing additional diversity in the mutant library (Wang, 1987), either through an increased size and mutation rate (Baker, 1985) or a lower selection strength (Prügel-Bennett and Shapiro, 1994).

IV. Rational Evolutionary Design

Dawinian evolution has proved a powerful optimization method for protein sequences. By utilizing the structure of the fitness landscape, tremendous improvements in both function and stability have been achieved. However, there are scenarios for which stepwise *in vitro* evolution may not be adequate. It is possible that achieving improved function requires more than single amino acid replacements. Discovering a triple or higher mutant in one experiment is a highly unlikely event. A related challenge is to discover enzymes with new or altered function (Shao and Arnold, 1996). Both processes occur in nature. The difficulty lies in reducing the time scale to an experimentally acceptable level.

In the previous section, theories applicable to each step of directed evolution were combined to provide insight into improving current methods. Section IV is devoted to descriptions of driving forces in evolution that have not been realized experimentrally. Aspects of the structure of the fitness landscape are described that have not been utilized in *in vitro* protein evolution. In Section IV.A, a connection is made between the tolerance, the designability, and the evolutionary potential of a protein. Section IV.B explores neutral theory and its potential to explore remote regions of sequence space, leading to the discovery of new structure and function. Finally, we focus on methods that reduce the time scale of evolution with the intention of inspiring future *in vitro* evolution experiments.

A. *Maximizing Evolvability*

Evolvability defines the potential for a protein to adapt to a new environment. In a sense, the molecular structure ''learns'' how to survive dynamic conditions. If a species has the ability to adapt to a changing environment, then it has a selective advantage over one that remains fixed (Maynard Smith, 1987). A bird that can learn how to eat fish has a selective advantage over one that cannot, if, for instance, the worm population suddenly declined. The fishing skill itself is not genetically passed to the offspring; however, the ability to learn is. This general principle is also applicable on the molecular level. If a protein has the ability to adapt quickly to a new environment, this provides a selective advantage in dynamic environments. Highly evolvable structures survive natural selection because of their tolerance for mutations (Li *et al.,* 1996). For example, viral coat proteins must repeatedly evade immune systems and therefore they have a higher tolerance for amino acid substitutions (Suzuki *et al.,* 1996a). In directed evolution experiments, the researcher changes the fitness landscape by screening for a new function, or a focused element of the natural fitness. Therefore, a protein that has previously adapted to changing environments will more quickly adjust to the new fitness requirements.

1. *Designability*

Structures are not evenly distributed in sequence space. A few structures dominate while most are represented by few or even no sequences (Li *et al.,* 1996). The fraction of sequence space that folds into a structure is defined as the designability of that structure. Designability is related to the collective tolerance of the residues. By definition, a structural scaffold that contains a large number of tolerant residues is designable. As discussed in Section II.C.1, directed evolution will be most successful on structurally tolerant proteins. Lattice models have been extensively used to demonstrate the relationship between designability, tolerance, and coupling. There is a tendency for mutations on a highly designable structure to be uncoupled (Li *et al.,* 1996). When the landscape is smooth, mutations can be accumulated to produce compound improvement. Therefore, this landscape is more likely to have single mutation paths that extend from the starting point to the region of higher fitness. This type of landscape—the landscape of designable structures—is good for directed evolution.

It has been postulated that stable structures are also designable (Li *et al.,* 1996). Mutations will not only alter the energy of the most stable structure, but also the energies of all other structures. When the energy gap between the most stable and the average structure is large, it is less likely that a mutation will cause the energy of a new structure to be lower than the stable structure (Tianna *et al.,* 1998; Broglia *et al.,* 1999; Buchler and Goldstein, 1999b). If the cumulative energetic impact of a series of point mutations is low enough not to disrupt the energy gap, then these mutants will fold to the same ground state, leading to structural tolerance. A correlation between stability and mutational robustness has also been observed for protein folding models (Govindarajan and Goldstein, 1997b; Bornberg-Bauer and Chan, 1999) and RNA secondary structure models (Ancel and Fontana, 1999).

Estimating tolerance is important in predicting the potential for directed evolution. As a method of determine the average tolerance over an entire protein sequence (designability) rather than tolerance at specific positions. Suzuki *et al.* (1996) compared the distribution of mutations in selected and unselected libraries. If the protein was tolerant to substitutions, the number of mutations in both libraries would be the same. The average number of mutations for the selected library is lowered by the fraction that is deleterious. Indeed, they found that the O-helix of Taq polymerase I (Taq Pol I) is significantly less tolerant than the β^3/β^4 region of HIV reverse transcriptase (HIV RT) (Fig. 17). They attribute this difference in survival rate versus number of mutations to

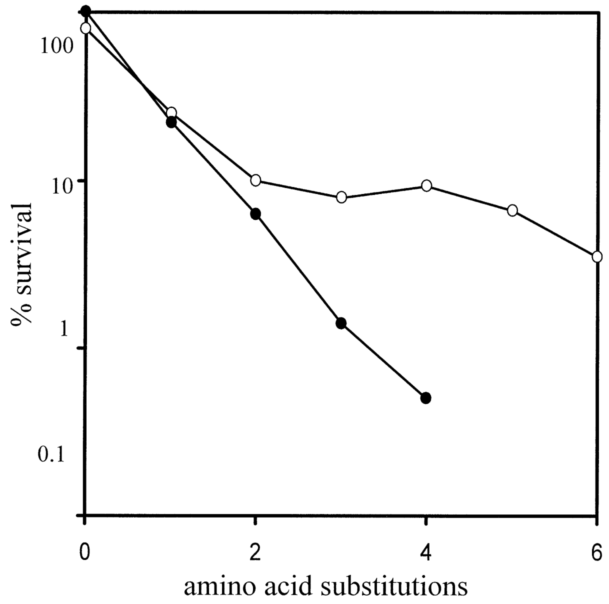


FIG. 17. The mutation rate (measured in number of amino acid substitutions) is plotted versus the log percentage of the mutant library that displays some fitness. The line for Taq Pol I (●) decreases rapidly, indicating an intolerant structure. Conversely, the line for HIV RT (○) stabilizes at high mutation rates, indicating a more tolerant structure. Reprinted from (Suzuki *et al.* 1996) with permission.

two possible mechanisms. First, from crystallographic studies, the β^3/β^4 region of HIV RT is known to be flexible and thus could accommodate more mutations. Second, HIV RT requires evolvability in order to evade host immunity, whereas the O-helix of Taq Pol I is involved in substrate binding and is highly conserved.

2. *The Evolution of Evolvability*

The optimal protein scaffold for an evolution experiment is tolerant, both structurally and functionally, to mutation. The overall tolerance (designability) of a structure is partially determined by the distribution of coupling interactions between residues. By minimizing the number of highly coupled residues, the ability for the protein to evolve in a dynamic environment improves considerably (Wilke and Martinez, 1999). Using a genetic model similar to a spin glass, Wagner (1996) showed that the pattern of coupling between residues evolves to act a buffer from the effect of mutations. Components that are highly conserved and undergo strong selection should show higher mutational robustness because it is more essential that they remain intact under the effects of mutation (Wagner and Stadler, 1999).

Several studies have shown that, under the influence of mutation, a population will evolve to highly tolerant regions of sequence space. Using a three-dimensional, maximally compact protein model, Govindarajan and Goldstein (1997b) showed that diffusion in structure space was more dependent on the selection pressure than on diffusion in sequence space. High selective pressure caused the structure to freeze while allowing the sequence to continue to evolve (Fig. 18). The less-tolerant sequences lose their fraction of the population more rapidly because there are more lethal mutations (Fontana and Schuster, 1987; Bornberg-Bauer and Chan, 1999; van Nimwegen *et al.,* 1999). This causes a flux to the regions of the network that have the most neutral neighbors, indicating that drift is the only driving force necessary for a population to achieve mutational tolerance.

Plasticity quantifies the ability of a sequence to be stable in multiple structural configurations or to perform multiple functions. A plastic protein is advantageous in a changing environment or if several simultaneous functions are required. The existence of a large ensemble of possible structures is costly to the fitness as the time spent in each single structure decreases (Ancel and Fontana, 1999). This is similar to the problem of an enzyme performing multiple tasks. As the number of required tasks increases, it becomes harder to optimize them all, thus creating landscape ruggedness. Plasticity causes a slowing of evolution when it is required that the population remain plastic, implying that there

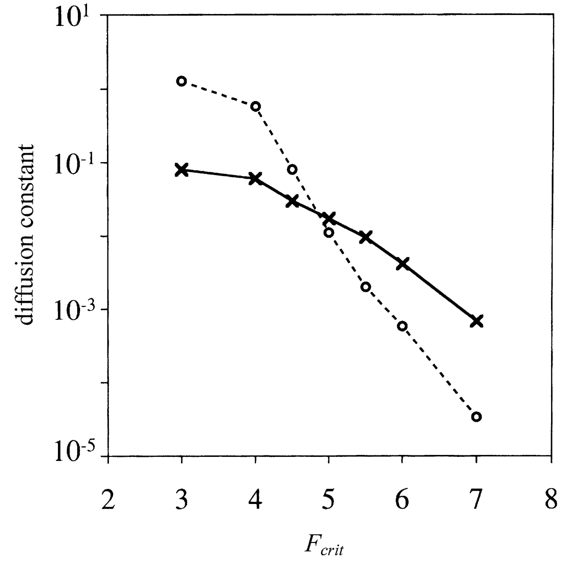


FIG. 18. Generalized diffusion constants for diffusion in sequence space (solid line) and diffusion in structure space (dotted line) as a function of the fitness cutoff F_{crit}. As F_{crit} increases, the selection strength increases. Diffusion in structure space is more dependent on F_{crit} than diffusion in sequence space. Reprinted from Govindarajan and Goldstein (1997b) with permission.

is no benefit for retaining plasticity in an *in vitro* evolution experiment. However, nature often requires that an enzyme be delicately controlled and perform multiple tasks. Because directed evolution is a driven process, selection acts to reduce the plasticity and, ultimately, fix the function and structure.

B. *Neutral Evolution*

Neutral evolution is based on the hypothesis that most mutations made on the molecular level do not alter the fitness of an organism (Kimura, 1983; Kimura, 1991). This is in contrast to adaptive, or Darwinian, evolution, which asserts that mutational changes that survive selection usually have a beneficial effect. There is empirical evidence supporting both theories (King and Jukes, 1969; Kimura, 1991; Eanes *et al.,* 1993; Hall, 1998). Recent efforts have focused on the relationship between neutral and adaptive evolution and insight has been provided into the mechanisms by which neutral evolution can drive adaptation. The neutral structure of a fitness landscape can aid the evolutionary search. Adaptive hill climbing on a rugged landscape tends to climb quickly to a local optimum (Fig. 19, see color insert). When neutrality is allowed, the population can diffuse across the hyperdimensional network and find paths to even higher peaks (Shuster *et al.,* 1994; Huynen, 1996; Huynen *et al.,* 1996; Newman and Engelhardt, 1998). During this period, neutral evolution allows for the buildup of multiple mutations so that a large mutational switch can be obtained. This process is stepwise, such that the population alternates between periods of random drift and rapid adaptive evolution (Schuster and Stadler, 1996).

The period when neutral drift dominates over adaptive evolution is called an epoch. If the fitness is plotted versus time, an epoch is a flat period showing no fitness improvement (Fig. 19). If a population becomes trapped at a suboptimal peak, it can remain trapped for a long time because it requires a specific series of mutations to escape the peak (Fontana *et al.,* 1989). The time that a population remains trapped is longer for larger peaks. Neutrality helps the evolving population to overcome fitness barriers so that the highest fitness achieved by evolution increases with the size of the neutral network, as more sequence space can be explored (Newman and Engelhardt, 1998).

A neutral network can arise out of conditions other than strict neutrality. The nearly neutral theory shows that if mutations only slightly disrupt the fitness, then some deviation in the sequence is allowed (Ohta, 1973; Ohta, 1997). This is closely related to the selection strength imposed on the population. In cases where there is a low selection strength, more deviations are allowed and the neutral network is larger. In addition, a neutral network can form due to noise in measuring the fitness (Levitan and Kauffman, 1995). Noise inherently limits the resolution at which selection can detect improvements so small changes are effectively, if not exactly neutral (Newman and Engelhardt, 1998).

To incorporate the benefits of neutrality into *in vitro* protein evolution experiments, it is necessary to understand the interactions between neutral networks of different structures and functions. In Section IV.B.1, we describe the structure of neutral networks based on measures developed primarily for RNA secondary structure landscapes. In Section IV.B.2, we focus on mutational transitions between networks. The behavior of evolving populations on neutral networks is described in Section IV.B.3, with the intention of driving future directed evolution experiments.

1. *Exploring Neutral Networks*

While a designable structure implies a large neutral network, the specific topology of the network is not defined. Designability only measures the volume of sequence space that folds into a structure, whereas a description of the connectivity is also required for a neutral network. One large network or many small networks can represent structures with the same designability, if the total sequence volume is equal. Very different sequences can exist on the same network if it extends through sequence space; if it is compact, however, very little space is covered. The total number and relative size of networks are also important: does one large network dominate or are there many equal-sized networks? Finally, the density of the network refers to the number of neutral paths that extend from a sequence point (Shuster *et al.,* 1994).

Newman and Engelhardt (1998) modified the *NK*-model to have degenerate energies, causing neutral networks to form. By adjusting the degree of degeneracy, the size of the networks could be tuned. The number of networks increases as A^N with the alphabet size A and sequence length N. The average size of the network increases as the landscape ruggedness increases (increasing K). The fraction of sequences in common networks goes to unity with large N, where a common network is defined as a network that contains a greater than average number of sequences. This implies that, for long sequences, a few common networks cover most of the space.

To discover new fitness peaks, the neutral network must be sufficiently extended, allowing neutral drift to effectively sample sequence space. A neutral network can be characterized by a mean fraction of neutral neighbors λ (Reidys *et al.,* 1997). If λ exceeds a threshold λ_c, then the network is connected and dense, making it more likely the network percolates through sequence space. If $\lambda < \lambda_c$, the networks are partitioned into components. Using random graph theory, the threshold is derived analytically as

$$\lambda_c = 1 - \sqrt[A-1]{A^{-1}}, \tag{32}$$

where A is the alphabet size (twenty for proteins). Networks with λ above this threshold are dense and connected, meaning that a single sequence undergoing neutral evolution can essentially explore the entire space. By definition, dense networks are more tolerant to mutations.

Although the existence of neutral networks has been well established for RNA, it is important to note the difference between RNA and protein landscapes. Most important, all RNA sequences fold into some structure, whereas it is likely that the vast majority of protein sequence space is devoid of well-defined structure. It may be possible to overcome this problem: because protein sequence space has a higher dimensionality, there are more chances to produce a connected network. In a preliminary study, Stadler and coworkers estimated the neutral structure of proteins by using an inverse folding algorithm to score the compatibility of a sequence with a structure (Babajide *et al.,* 1997). The z-score is a measure of the difference between E_i, the energy of sequence i folded into the target structure and $<E>$, the average energy of sequences folding into the structure:

$$z(i) = \frac{E_i - <E>}{\sigma(E)}, \tag{33}$$

where σ indicates the standard deviation. A sequence is considered a member of a structure's neutral network if the z-score is lower than a critical threshold, z_c. Surprisingly, it was found that the average length of the neutral walk was 90% to 95% the number of residues in the protein, providing preliminary evidence that protein neutral networks have the potential to span sequence space. When the alphabet was reduced to only four amino acids, the size of the networks decreased; however, they remained expansive, comprising roughly 80% of the residues. This implies that, despite the amino-acid substitution restrictions imposed on directed evolution by the genetic code, extended and dense neutral networks can exist. These results are consistent with experimental observations that very diverse sequences can fold into similar structures (Aronson *et al.,* 1994).

2. *Discovering Novel Structure and Function*

Sequence space is a patchwork of neutral networks. Because this space is of high dimensionality, the majority of sequence points lie on the surface of each network. This is due to the topology of the space: When the dimensionality is large (in N), the volume approaches the surface area. This implies that connected neutral networks contact each other frequently, like interwoven sponges. Assuming that a sequence that evolved naturally has a high designability, the neutral network of the structure will be correspondingly large with extensive connecting paths between clusters. The possibilities for innovation are endless. Because the neutral networks from different structures are interwoven, a population drifting on the neutral network frequently produces mutants that fall off the network, discovering new networks in the process (Huynen *et al.,* 1996). Through this mechanism, neutral evolution fulfills an important role in adaptation by enabling exploration of new structure and function space.

Most of the studies of the ability of neutral evolution to discover new phenotypes have been carried out on RNA secondary structure models. Here, the neutral network is defined as a connected region of sequence space that folds into the same secondary structure. To apply the results to protein evolution, the secondary structure neutral network can be considered as a connected region of similar functions or protein tertiary structures. Huynen (1996) observed that a walk on a neutral net, as opposed to an unconstrained walk in sequence space, increases the likelihood of finding new structures through a single mutant. In general, when the potential for neutrality is decreased, the ability to find new structures also decreases. In all cases, the number of different structures observed increased linearly with the length of the walk. Neutral evolution samples a nearly limitless number of structures, thereby enhancing adaptive evolution.

Compared to the diameter of sequence space, the correlation length of structure space for RNA folding is relatively small (Fontana *et al.,* 1993). A small correlation length implies that a small sphere around any sequence can sample all possible secondary structures. The ability to sample many structures from any sequence point is a property of the fitness landscape referred to as ''shape space covering.'' Equation (32) predicts shape space covering of structures when the connectivity is greater than λ_c (Reidys *et al.,* 1997). The radius of the covering sphere r_{cov} is defined as

$$r_{cov} = \min \{h | B_h \geq S_N\} \tag{34}$$

where B_h is the number of sequences in the sphere of radius h, and S_N is the total number of secondary structures given a sequence of length N (Schuster and Stadler, 1996). The covering radius is small compared with the entire sequence space. However, it contains an evolutionarily representative proportion of structure.

The concept of shape space covering was first extended from structure to function in order to study antibodies. Each antibody evolves to cover a region of antigen shape space, where shape space is now defined as the ability to bind the antigen (Perelson and Oster, 1979; Perelson, 1990; Hightower *et al.,* 1995). Each dimension in this space represents one of the requirements for binding antigens, for example, hydrophobicity, shape complement, hydrogen bonding, and electrostatic interactions. The specificity of an antibody for an antigen is never perfect, thus, each antibody covers a small volume of antigen space. The volumes can overlap, creating a redundancy in the ability of different antibodies to bind to the same antigen. The optimal antibody library balances the need for redundancy to guarantee binding to all antigens with the limited number of antibodies that the immune system can produce. Using heuristic arguments, it is estimated that the total number of antibodies required to sample this space is roughly 10^8, or the same number of antibodies that can be produced by the human immune system (Perelson and Oster, 1979). By simulating the immune system using a genetic algorithm, it was demonstrated that the antibody gene library repeatedly converged on the optimal antigen space covering (Hightower *et al.,* 1995).

It is necessary to develop a theory to characterize movements through antigen (or functional) space that result from the property of shape space covering. Returning conceptually to RNA, the behavior of transitions between networks is related to the structural rearrangement that is required (Fontana and Schuster, 1998). Transitions are either continuous or discontinuous, based on the required state of other positions in the structure. Continuous transitions require a small rearrangement, such as the addition or removal of a base pair in RNA. Conversely, discontinuous transitions require much larger rearrangements, such as a shift in the base pairing. The epoch periods in neutral evolution are usually dominated by a series of continuous transitions, whereas the larger, punctuated changes correlate with a discontinuous transition. A discontinuous transition can only be activated if all the remaining positions in the sequence are in the correct state to facilitate the change. The epoch periods represent the time required for neutral drift to randomly produce the correct combination of amino acids or base pairs to cause a punctuated switch. To reduce the time spent in epochs, it is important to characterize the ability to take a single step walk through shape space, similar to the necessity of a connected network in sequence space, as postulated by Maynard Smith (Fig. 20, see color insert).

These concepts are grossly related to the ability of *in vitro* evolution to discover new functions. Directing the evolution of new function is difficult because it is not known how far a jump in sequence space is required (Arnold and Wintrode, 1999). If the new function can be generated by a few mutations, then mutagenesis and screening may be all that is required. However, if a large number of coordinated mutations is required, the probability of this occurring in a reasonably sized mutant library is negligible. The alternate means by which this may be achieved is to develop methods to discover the connected pathway in shape (or now catalytic) space that allow regions of new function to be explored.

The range of enzyme function can be condensed into catalytic task space, where each point represents a catalytic task and an enzyme covers a portion of the space corresponding to its repertoire of functions (Kauffman, 1992; Kauffman, 1993). In addition to the shape features, the axes include chemical and physical aspects relevant to catalysis. This construction of catalytic space was motivated by the discovery of antibody catalysis through binding of the transition state of a reaction (Lienhrd, 1973; Pollack *et al.,* 1986; Benkovic *et al.,* 1990; Driggers and Schultz, 1996; Xiu *et al.,* 1996). In this sense, the catalytic space of the enzyme is the binding of the transition state; similar reactions bind to similar transition states. Because catalytic specificity is never perfect, any enzyme can catalyze a range of reactions, represented by a ball in catalytic space. This indicates that only a finite number of enzymes may be necessary to cover all simple catalytic tasks.

Mutations can accomplish two movements in catalytic space: first, they can increase the specificity and cause a shrinkage of the ball, and, second, they can cause the ball to shift or expand, thus changing the ensemble of functions that the enzyme can perform. Both concepts are crucial for *in vitro* evolution studies because the first task involves improving an existing function and the second task is the gradual stepwise discovery of novel function. To understand transitions in catalytic space, it is necessary to define the transitions induced by mutations in proteins, as was done for RNA. To some extent, these transitions have been identified for structural transitions. It has been found that many amino acid substitutions lead to a continuous change in the protein structure (Chothia and Lesk, 1986). Most changes permuted the local structure or shifted the relative position of secondary structure elements. It has been observed experimentally that different structures can exist close to one another in sequence space, implying shape space covering (Jones *et al.,* 1996; Pawlowski *et al.,* 1996). Transitions between networks have also be observed as punctuated jumps in structure from single amino acid changes (Cordes *et al.,* 1999; Glykos *et al.,* 1999).

In addition to understanding structural transitions, it is important to understand the properties that make a functional transition continuous or discontinuous. The degree of continuity of a functional transition is related to the overlap of the two functions in sequence space. Urabe and co-workers studied the overlap in sequence space between catalase activity, peroxidase activity, and thermostability for catalase I (Trakulnaleamsai *et al.,* 1995). A positive correlation between the two catalytic activities was observed, indicating that the structure of the landscape is similar for both. The source of this overlap is that both activities involve the same catalytic site and share a common intermediate. Correlation between catalase/peroxidase activity and thermostability was much weaker, indicating less overlap in sequence space, implying that transitions between similar functions are continuous. The best method to evolve sequences with new function is to evolve through a set of gradual, continuous changes. Here, an essential problem is defined in developing methods to recognize the intermediates on the path between one function and the next. To harness neutral evolution, a theory is required that differentiates selection for the smooth (continuous) transitions with the punctuated (discontinuous) transitions.

3. Is In Vitro Neutral Evolution Practical?

The difficulty in utilizing neutral evolution is the inherent dependence on stochastic movement, without its showing a gain in fitness. This process is not easily exploited by directed evolution. However, it may be possible to apply the principles of neutral evolution to an otherwise adaptive search. The time scale for neutral evolution to produce adaptive improvements is the time spent in an epoch. The lengths of the epochs were found to increase with increasing fitness, indicating that it becomes more difficult to utilize neutral evolution as the sequence becomes more optimized (Newman and Engelhardt, 1998). To accomplish the goal of reducing the epoch time, a theory is required that defines the barriers for punctuated changes to occur. Reducing these barriers by tuning the evolutionary parameters will expedite the neutral search.

The epoch time is dominated by the time required for neutral drift to align the amino acids in their correct state in order to facilitate an adaptive change. Lattice model studies imply that it may require more than a single mutation to make a transition to a new neutral network (Lipman and Wilbur, 1991; Bornberg-Bauer, 1997). Unfortunately, the requirement that multiple mutations be made to observe a jump in fitness requires more time than overcoming a large fitness barrier with fewer mutations (van Nimwegen and Crutchfield, 1999b). Empirically, the scaling function for the required time to cross a barrier is given by Eq. (35)

$$<t>\propto \frac{1}{w!\ Mp_m}\left(\frac{\log(s)}{p_m}\right)^{w-1}, \tag{35}$$

where w is the barrier width (number of required mutations), s is the barrier depth, M is the population size, and p_m is the mutation rate. The crossing time scales as a power law in both $\log(s)$ and p_m. The scaling is most rapid with w, indicating that the barrier width dominates the discovery time. If a fitness barrier crossing is required, then the sequence is most likely to take the path requiring the fewest mutations. In a related study, Zhang (1997) studied the behavior of the quasispecies with finite populations. Stochastic noise moves the sequence off a local optimum, allowing the search of more space. The drift distance, d off a local optimum as a function of time t was found for a finite population

$$d(t) = \frac{\ln M + \ln t}{|\ln p_m|}. \tag{36}$$

The expected time to find a better peak is $t_d = 1/(Mp_m^d)$, similar to a prediction by Gillespie (1984). Both Eq. (35) and (36) predict that it is more important to have a large mutation rate rather than a large population size to accelerate an adaptive change.

Van Nimwegen and Crutchfield (1999a) have constructed a theory for the optimization of evolutionary searches involving epochal dynamics. They showed that the destabilization of the epochs due to fluctuations in the finite population occurs near the optimal mutation rate and population size. Under these conditions, the epoch time is only constrained by the diffusion of the population to a neutral network boundary. Often the optimal parameters are very close to the region in which destabilization is an important effect. This emphasizes that, to utilize neutral evolution, it is important to tune the evolutionary parameters (such as mutation rate and population size) so that the time spent in an epoch is minimized without destabilizing the search.

Simulations of RNA secondary structure landscapes provide insight into the necessary mutation rate to drive adaptation. Huynen *et al.* (1996) found that the ability of a population to adapt is determined by the error threshold of the fitness and not the sequence. Indeed, they found that any mutation rate greater than zero will cause the population to drift on the neutral network [The error threshold on landscapes with high neutrality approaches zero (Derrida and Peliti, 1991).] A second, higher mutation threshold causes the fitness information to be lost. To accelerate the diffusion of the population on the neutral network, it is necessary to be above the sequence error threshold and as close to the fitness error threshold as possible. Under these criteria, the population will diffuse rapidly without losing fitness information. On a flat landscape, the diffusion constant D_0 for a population of M sequences of length N can be approximated by Eq. (37).

$$D_0 \approx \frac{5aNp_m}{3 + 4Mp_m}, \tag{37}$$

where a is the replication rate and p_m is the mutation rate (Huynen *et al.,* 1996). For small mutation rates, the diffusion constant can be approximated by $D = D_0\lambda$, where λ is the average fraction of neutral mutants for the dominant structure. By using the Kimura's relation $k = ap_mN\lambda$ (where k is the fixation rate per generation and can be interpreted as the rate of entropy loss) (Kimura, 1983), the result

$$D = \frac{6k}{3 + 4Mp_m} \tag{38}$$

is shown to be in good agreement with the computational results on theoretical RNA secondary structure landscape (Fig. 21). Initially, the diffusion constant increases rapidly with mutation rate and then levels off. Beyond this point, little speed is gained by increasing the mutation rate. This transition occurs at a mutation rate that is obtainble in directed evolution experiments.

Tachida (1991) characterized the behavior of a population on a neutral network based on the parameter $4M\sigma$, where M is the population size and σ is the standard deviation of the effects of mutations on the fitness. The parameter σ is related to the tolerance of a residue in a protein. If σ is large, then the effect of mutations is large. Three types of behavior are identified in this model. In the case $4M\sigma < 0.2$, the

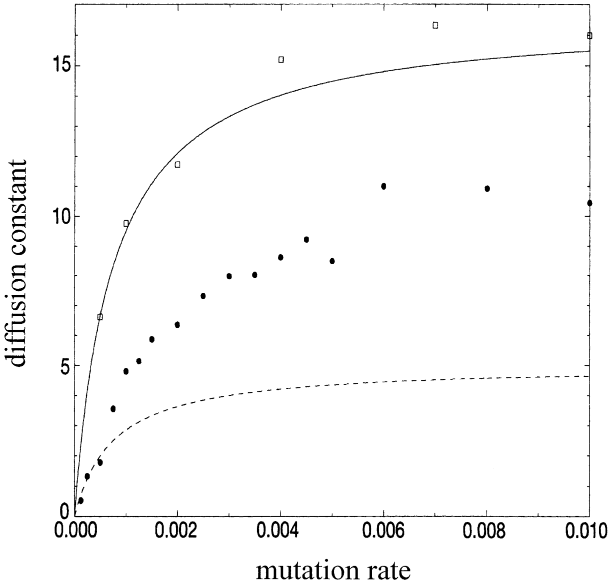


FIG. 21. The diffusion constant on a neutral network D_0 is plotted versus the mutation rate p_m. The simulations (●) are for RNA sequence length $N = 76$ and population size $M = 1000$ and are allowed to equilibirate before the statistics are taken. The solid line is the theoretical D_0 and (□) are flow-reactor simulations for a flat landscape. The dotted line is calculated for $D_0\lambda$, where $\lambda = 0.3$ is the estimated fraction of neutral mutants. Reprinted from Huynen *et al.* (1996) with permission. Copyright (1996) National Academy of Sciences, USA.

substitutions behave neutrally. In the intermediate case $0.2 < 4M\sigma < 3 - 5$, the behavior is nearly neutral; both neutral and advantageous mutations are fixed. In the case $4M\sigma > 3 - 5$, the initial rise in fitness is rapid, after which the approach to equilibrium is slow. Only advantageous mutations and not neutral mutations occur in the initial burst phase of adaptation. The residues in a protein have a variety of σ values (related to the tolerance of that residue). Each of these positions behaves differently in evolution, according to one of the regimes described above. The behavior of evolution can then be effectively tuned between the neutral case and the adaptive case by targeting mutagenesis at different regions of the gene.

In accelerating neutral evolution, it is clearly important to increase the mutation rate to decrease the epoch times. Equations (35) and (36) predict that increasing the mutation rate is more productive than increasing the screening effort in obtaining a multi-mutational switch. The increased mutation rate increases the diffusion of a population on a neutral network, allowing a broad sampling of sequence space. However, as discussed in Section II.A.3, a high muation rate has a negative overall effect because of the screening costs and increased appearance of stop codons. By targeting the regions of the sequence based on σ, the neutrality of the search can be tuned. The value of σ for each residue can be estimated using methods described in Section III.C.2. A smooth (or neutral) landscape can accept more mutations, due to the larger correlation length (Section II.A.3). Thus, the ability for neutral evolution to make an adaptive change has been accelerated.

V. Summary and Conclusions

Directed evolution uses a combination of powerful search techniques to generate proteins with improved properties. Part of the success is due to the stochastic element of random mutagenesis; improvements can be made without a detailed description of the complex interactions that constitute function or stability. However, optimization is not a conglomeration of random processes. Rather, it requires both knowledge of the system that is being optimized and a logical series of techniques that best explores the pathways of evolution (Eigen *et al.,* 1988). The weighing of parameters associated with mutation, recombination, and screening to achieve the maximum fitness improvement is the beginning of rational evolutionary design.

The optimal mutation rate is strongly influenced by the finite number of mutants that can be screened. A smooth fitness landscape implies that many mutations can be accumulated without disrupting the fitness.

This has the effect of lowering the required library size to sample a higher mutation rate. As the sequence ascends the fitness landscape, the optimal mutation rate decreases as the probability of discovering improved mutations also decreases. Highly coupled regions require that many mutations be simultaneously made to generate a positive mutant. Therefore, positive mutations are discovered at uncoupled positions as the fitness of the parent increases.

The benefit of recombination is twofold: it combines good mutations and searches more sequence space in a meaningful way. Recombination is most beneficial when the number of mutants that can be screened is limited and the landscape is of an intermediate ruggedness. The structure of schema in proteins leads to the conclusion that many cut points are required. The number of parents and their sequence identity are determined by the balance between exploration and exploitation. Many disparate parents can explore more space, but at the risk of losing information.

The required screening effort is relatd to the number of uphill paths, which decreases more rapidly for rugged landscapes. Noise in the fitness measurements causes a dramatic increase in the required mutant library size, thus implying a smaller optimal mutation rate. Because of strict limitations on the number of mutants that can be screened, there is motivation to optimize the content of the mutant library. By restricting mutations to regions of the gene that are expected to show improvement, a greater return can be made with the same number of mutants. Initial studies with subtilisin E have shown that structurally tolerant positions tend to be where positive activity mutants are made during directed evolution. Mutant fitness information is produced by the screening step that has the potential to provide insight into the structure of the fitness landscape, thus aiding the setting of experimental parameters. By analyzing the mutant fitness distribution and targeting specific regions of the sequence, *in vitro* evolution can be accelerated. However, when expediting the search, there is a trade-off between rapid improvement and the quality of the long-term solution.

The benefit of neutrality has yet to be captured with *in vitro* protein evolution. Neutral theory predicts the punctuated emergence of novel structure and function, however, with current methods, the required time scale is not feasible. Utilizing neutral evolution to accelerate the discovery of new functional and structural solutions requires a theory that predicts the behavior of mutational pathways between networks. Because the transition from neutral to adaptive evolution requires a multi-mutational switch, increasing the mutation rate decreases the time required for a punctuated change to occur. By limiting the search to the less coupled region of the sequence (smooth portion of the fitness landscape), the required larger mutation rate can be tolerated. Advances in directed evolution will be achieved when the driving forces behind such processes are captured theoretically and applied experimentally. Like paths in sequence space, there are unlimited possibilities leading everywhere.

Acknowledgments

The authors express gratitude to Frances Arnold who has provided wonderful insight and guidance throughout the preparation of this manuscript. The authors also thank Steven Mayo and Gillian Pierce for critical readings and useful discussions. CAV is supported under a National Science Foundation Graduate Fellowship.

References

Abkevich, V. I., Gutin, A. M., and Shakhnovich, E. I. (1995). *J. Mol. Biol.* **252,** 460–471.

Aita, T., and Husimi, Y. (1996). *J. Theor. Biol.* **182,** 469–485.

Aita, T., and Husimi, Y. (1998a). *J. Theor. Biol.* **193,** 383–405.

Aita, T., and Husimi, Y. (1998b). *J. Theor. Biol.* **191,** 377–390.

Amitrano, C., Peliti, L., and Saber, M. (1989). *J. Mol. Evol.* **29,** 513–525.

Ancel, L. W., and Fontana, W. (2000). *J. of Exp. Zoology,* in press.

Anderson, P. W. (1983). *Proc. Natl. Acad. Sci. USA.* **80,** 3386–3390.

Arkin, A. P., and Youvan, D. C. (1992). *Proc. Natl. Acad. Sci. USA.* **89,** 7811–7815.

Arnold, F. H., and Wintrode, P. L. (1999). In ''Encyclopedia of Bioprocess Technology: Fermentation, Biocatalysis, and Bioseparation'' (M. C. Flickinger and S. W. Drew, eds.), Vol. 2, pp. 971–987. John Wiley & Sons, Inc., New York.

Aronson, H.-E. G., Royer, W. E., and Hendrickson, W. A. (1994). *Protein Science.* **3,** 1706–1711.

Babajide, A., Hofacker, I. L., Sippl, M. J., and Stadler, P. F. (1997). *Folding & Design.* **2,** 261–269.

Bäk, T. (1992). *Parallel Problem Solving from Nature 2, Brussels.* 85–94.

Bäk, T. (1993). *Proceedings of the Fifth International Conference on Genetic Algorithms,* Urbana-Champaign. 2–8.

Baker, J. E. (1985). *Proceedings of the First International Conference on Genetic Algorithms and Their Applications,* Pittsburgh. 101–111.

Benkovic, S. J., Adams, J. A., C. L. Borders, J., Janda, K. D., and Lerner, R. A. (1990). *Science.* **250,** 1135–1139.

Bergman, A., and Feldman, M. W. (1992). *Physica D.* **56,** 57–67.

Bogarad, L. D., and Deem, M. W. (1999). *Proc. Natl. Acad. Sci. USA.* **96,** 2591–2595.

Bonhoeffer, S., and Stadler, P. F. (1993). *J. Theor. Biol.* **164,** 359–372.

Bonnaz, D., and Koch, A. J. (1998). *J. Phys. A: Math. Gen.* **31,** 417–429.

Bornberg-Bauer, E. (1997). *Biophys. J.* **73,** 2393–2403.

Bornberg-Bauer, E., and Chan, H. S. (1999). *Proc. Natl. Acad. Sci. USA.* **96,** 10689–10694.

Bornholdt, S. (1998). *Phys. Rev. E.* **57,** 3853–3860.

Borstnik, B., Pumpernik, D., and Hofacker, G. L. (1987). *J. Theor. Biol.* **125,** 249–268.

Brady, R. M. (1985). *Nature.* **317,** 804–806.

Broglia, R. A., Tiana, G., Roman, H. E., Vigezzi, E., and Shakhnovich, E. (1999). *Phys. Rev. Lett.* **82,** 4727–4730.

Brown, B. M., and Sauer, R. T. (1999). *Proc. Natl. Acad. Sci. USA.* **96,** 1983–1988.

Bryngelson, J. D., and Wolynes, P. G. (1987). *Proc. Natl. Acad. Sci. USA.* **84,** 7524–7528.

Buchler, N. E. G., and Goldstein, R. A. (1999a). *Proteins: Struct, Funct, and Gen.* **34,** 113–124.

Buchler, N. E. G., and Goldstein, R. A. (1999b). *J. Chem. Phys.* **111,** 6599–6609.

Chen, K., and Arnold, F. H. (1993). *Proc. Natl. Acad. Sci. USA.* **90,** 5618–5622.

Chien, N. C., Roberts, V. A., Giusti, A. M., Scharff, M. D., and Getzoff, E. D. (1989). *Proc. Natl. Acad. Sci. USA.* **86,** 5532–5536.

Chothia, C., and Lesk, A. M. (1986). *EMBO J.* **5,** 823–826.

Coker, P., and Winter, C. (1997). *Fourth European Conference on Artificial Life.* 163–169.

Cordes, M. H. J., Walsh, N. P., McKnight, C. J., and Sauer, R. T. (1999). *Science.* **284,** 325–327.

Craighurst, R., and Martin, W. (1995). *Proceedings of the Sixth International Conference on Genetic Algorithms,* Pittsburgh. 130–135.

Crameri, A., Raillard, S.-A., Bermudez, E., and Stemmer, W.P.C. (1998). *Nature.* **391,** 288–291.

Dahiyat, B. I., and Mayo, S. L. (1997). *Science.* **278,** 82–87.

Darwin, C. (1859). ''The Origin of Species.'' Murray, London.

Daugherty, P. S., Chen, G., Olsen, M. J., Iverson, B. L., and Georgiou, G. (1998). *Protein Engineering.* **11,** 825–832.

Daugherty, P. S., Chen G., Iverson, B. L., and Georgiou, G. (2000). *Proc. Natl. Acad. Sci. USA.* **97,** 2029–2034.

De Jong, K. A., and Spears, W. M. (1990). *Parallel Problem Solving from Nature I.* Dortmund. 38–47.

Delagrave, S., Goldman, E. R., and Youvan, D. C. (1995). *Protein Engineering.* **8,** 237–242.

Delagrave, S., and Youvan, D. C. (1993). *Bio/Technology.* **11,** 1548–1552.

Derrida, B. (1980). *Phys. Rev. Lett.* **45,** 79–83.

Derrida, B. (1981). *Phys. Rev. B.* **24,** 2613–2626.

Derrida, B., and Peliti, L. (1991). *Bull. Math. Biol.* **53,** 335–381.

Dewey, T. G., and Donne, M. D. (1998). *J. Theor. Biol.* **193,** 593–599.

Dill, K. A., Bromberg, S., Yue, K., Fiebig, K. M., Yee, D. P., Thomas, P. D., and Chen, H. S. (1995). *Protein Science.* **4,** 561–602.

Driggers, E. M., and Schultz, P. G. (1996). *Advances in Protein Chemistry.* **49,** 261–287.

Dube, D. K., Black, M. E., Munir, K. M., and Loeb, L. A. (1993). *Gene.* **137,** 41–47.

Dunbrack, R. L., and Karplus, M. (1994). *Nature Struct. Biol.* **1,** 334–340.

Eanes, W. F., Kirchner, M., and Yoon, J. (1993). *Proc. Natl. Acad. Sci. USA.* **90,** 7475–7479.

Eiben, A. E., van Kemenade, C. H. M., and Kok, J. N. (1995). *Advances in Artificial Life.* **929,** 934–945.

Eiben, M. E., and Schippers, C. A. (1996). *Parallel Problem Solving from Nature IV.* Berlin. 319–328.

Eigen, M. (1971). *Naturwissenschaften.* **58,** 465–523.

Eigen, M., McCaskill, J., and Schuster, P. (1989). *Adv. Chem. Phys.* **75,** 149–263.

Eigen, M., McCaskill, J., and Shuster, P. (1988). *J. Phys. Chem.* **92,** 6881–6891.

Eschelman, L. J., Caruana, R. A., and Schaffer, J. D. (1989). *Proceedings of the Third International Conference on Genetic Algorithms,* George Mason University. 10–19.

Eshelman, L. J., and Schaffer, J. D. (1991). *Proceedings of the Fourth International Conference on Genetic Algorithms,* San Diego. 115–121.

Eshelman, L. J., and Schaffer, J. D. (1995). In ''Foundations of Genetic Algorithms 3'' (L. D. Whitley and M. D. Vose, eds.), pp. 299–313. Morgan Kaufmann, San Francisco.

Fischer, K. H., and Hertz, J. A. (1991). ''Spin Glasses.'' Cambridge University Press, Cambridge.

Flyvbjerg, H., and Lautrup, B. (1992). *Phys. Rev. A.* **46,** 6714–6723.

Fogarty, T. C. (1989). *Proceedings of the Third International Conference of Genetic Algorithms,* George Mason University. 104–109.

Fontana, W., Schnabl, W., and Shuster, P. (1989). *Phys. Rev. A.* **40,** 3301–3321.

Fontana, W., and Schuster, P. (1998). *J. Theor. Biol.* **194,** 491–511.

Fontana, W., and Shuster, P. (1987). *Biophys. Chem.* **26,** 123–147.

Fontana, W., and Shuster, P. (1998). *Science.* **280,** 1451–1455.

Fontana, W., Stadler, P. F., Bornberg-Bauer, E. G., Griesmacher, T., Hofacker, I. L., Tacker, M., Tarazona, P., Weinberger, E. D., and Schuster, P. (1993). *Phys. Rev. E.* **47,** 2083–2099.

Forrest, S. (1993). *Science.* **261,** 872–878.

Forrest, S., and Mitchell, M. (1993). In ''Foundations of Genetic Algorithms 2'' (L. D. Whitley, ed.), pp. 109–126. Morgan Kaufmann, San Mateo, California.

Frauenfelder, H., Sligar, S. G., and Wolynes, P. G. (1991). *Science.* **254,** 1598–1063.

García-Pelayo, R., and Stadler, P. F. (1997). *Physica D.* **107,** 240–254.

Geysen, H. M., Rodda, S. J., Mason, T. J., Tribbick, G., and Schoofs, P. G. (1987). *J. Immun. Methods.* **102,** 259–274.

Gillespie, J. H., (1984). *Evolution.* **38,** 1116–1129.

Glykos, N. M., Cesareni, G., and Kokkinidis, M. (1999). *Structure.* **7,** 597–603.

Goldman, N., Thorne, J. L., and Jones, D. T. (1998). *Genetics.* **149,** 445–458.

Govindarajan, S., and Goldstein, R. (1997a). *Biopolymers.* **42,** 427–438.

Govindarajan, S., and Goldstein, R. A. (1997b). *Proteins.* **29,** 461–466.

Grefenstette, J. J. (1995). In ''Foundations of Genetic Algorithms 3'' (L. D. Whitley and M. D. Vose, eds.), pp. 139–161, Morgan Kaufmann, San Francisco.

Hall, B. G. (1998). *Genetica.* **102/103,** 109–125.

Hamming, R. W. (1950). *Bell Syst. Tech. J.* **29,** 147–160.

Hawrani, A. S. E., Moreton, K. M., Sessions, R. B., Clarke, A. R., and Holbrook, J. J. (1994). *TIBTECH.* **12,** 207-211.

Hellinga, H. W., Wynn, R., and Richards, F. M. (1992). *Biochemistry.* **31,** 11203–11209.

Hesser, J., and Männer, R. (1990). *Parallel Problem Solving from Nature I, Dortmund* 23–28.

Hightower, R. R., Forrest, S., and Perelson, A. S. (1995). *Proceedings of the Sixth International Conference on Genetic Algorithms,* Pittsburgh. 344–350.

Holland, J. (1975). ''Adaptation in Natural and Artificial Systems.'' The University of Michigan Press, Ann Arbor, MI.

Hordijk, W., and Manderick, B. (1995). *Advances in Artificial Life.* **929,** 908–919.

Houghten, R. A., Pinilla, C., Blondelle, S. E., Appel, J. R., Dooley, C. T., and Ceurvo, J. H. (1991). *Nature.* **354,** 84–86.

Huynen, M. A. (1996). *J. Mol. Evol.* **43,** 165–169.

Huynen, M. A., Stadler, P. F., and Fontana, W. (1996). *Proc. Natl. Acad. Sci. USA.* **93,** 397–401.

Irvine, D., Tuerk, C., and Gold, L. (1991). *J. Mol. Biol.* **222,** 739–761.

Jencks, W. P. (1981). *Proc. Natl. Acad. Sci. USA.* **78,** 4046–4050.

Jones, D. T., Moody, C. M., Uppenbrink, J., Viles, J. H., Doyle, P. M., Harris, C. J., Pearl, L. H., Sadler, P. J., and Thornton, J. M. (1996). *Proteins.* **24,** 502–513.

Jones, T., and Forrest, S. (1995). *Proceedings of the Sixth International Conference on Genetic Algorithms,* Pittsburgh. 184–192.

Joo, H., Lin, Z., and Arnold, F. H. (1999). *Nature.* **399,** 670–673.

Kamtekar, S., Schiffer, J. M., Xiong, H., Babik, J. M., and Hecht, M. H. (1993). *Science.* **262,** 1680–1685.

Kauffman, S. (1993). ''The Origins of Order.'' Oxford University Press, Oxford.

Kauffman, S., and Levin, S. (1987). *J. Theor. Biol.* **128,** 11–45.

Kauffman, S. A. (1992). *J. Theor. Biol.* **157,** 1–7.

Kauffman, S. A., and Macready, W. G. (1995). *J. Theor. Biol.* **173,** 427–440.

Kauffman, S. A., and Weinberger, E. D. (1989). *J. Theor. Biol.* **141,** 211–245.

Kimura, M. (1983). ''The Neutral Theory of Molecular Evolution.'' Cambridge University Press, Cambridge.

Kimura, M. (1991). *Jpn. J. Genet.* **66,** 367–386.

King, J. L., and Jukes, T. H. (1969). *Science.* **164,** 788–798.

Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P. (1983). *Science.* **220,** 671–680.

Knowles, J. R. (1991). *Nature.* **350,** 121–124.

Koehl, P., and Delaru, M. (1996). *Curr. Opin, Struct. Biol.* **6,** 222–226.

Kuchner, O., and Arnold, F. H. (1997). *TIBTECH.* **15,** 523–530.

Lee, C. (1994). *J. Mol. Biol.* **236,** 918–939.

Levinthal, C. (1969). *Mossbauer Spectroscopy in Biological Systems,* Monticelli, Illinois. 22.

Levitan, B. (1997). *In* ''Annual Reports in Combinatorial Chemistry and Molecular Diversity'' (W. H. Moos, M. R. Pavia, B. K. Kay, and A. D. Ellington, eds), Vol. 1, pp. 95–152. Escom, Leiden.

Levitan, B. (1998). *J. Mol. Biol.* **277,** 893–916.

Levitan, B., and Kauffman, S. (1995). *Molecular Diversity.* **1,** 53–68.

Li, H., Helling, R., Tang, C., and Wingreen, N. (1996). *Science.* **273,** 666–669.

Li, W.-H., and Graur, D. (1991). ''Fundamentals of Molecular Evolution.'' Sinauer Associates, Sunderland.

LiCata, V. J., and Ackers, G. K. (1995). *Biochemistry.* **34,** 3133–3139.

Lienhrd, G. E. (1973). *Science.* **180,** 149–154.

Lipman, D. J., and Wilbur, J. (1991). *P. Roy. Soc. Lond. B.* **245,** 7–11.

Macken, C. A., Hagan, P. S., and Perelson, A. S. (1991). *SIAM J. Appl. Math.* **51,** 799–827.

Macken, C. A., and Perelson, A. S. (1989). *Proc. Natl. Acad. Sci. USA.* **86,** 6191–6195.

Macken, C. A., and Perelson, A. S. (1991). In ''Molecular Evolution on Rugged Landscapes'' (A. S. Perelson and S. A. Kauffman, eds), Vol. IX, pp. 93–118. Addison-Wesley.

Macken, C. A., and Stadler, P. F., (1993). In ''Lectures in Complex Systems'' (L. Nadel and D. L. Stein, eds.), Vol. VI, pp. 43–86, Addison-Wesley.

Mandecki, W., Chen, Y.-C. J., and Grihalde, N. (1995). *J. Theor. Biol.* **176,** 523–530.

Manderick, B., de Weger, M., and Spiessens, P. (1991). *Proceedings of the Fourth International Conference on Genetic Algorithms,* San Diego. 143–150.

Martinez, H. M. (1984). *Nucl. Acid. Res.* **12,** 323–334.

Matsuura, T., Yomo, T., Trakulnaleamsai, S., Ohashi, Y., Yamamoto, K., and Urabe, I. (1998). *Protein Engineering.* **11,** 789–795.

Maynard Smith, J. (1970). *Nature.* **225,** 563–564.

Maynard Smith, J. (1987). *Nature.* **329,** 761–762.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). *J. Chem. Phys.* **21,** 1087–1092.

Michod, R. E. (1999). ''Darwinian Dynamics.'' Princeton University Press, Princeton.

Mitchell, M. (1998). ''An Introduction to Genetic Algorithms.'' The MIT Press, Cambridge.

Miyazaki, K., and Arnold, F. H. (1999). *J. Mol. Evol.* **49,** 716–720.

Miyazawa, S., and Jernigan, R. (1985). *Macromolecules.* **18,** 534.

Montoya F., and Dubois J.-M. (1993). *Europhys. Lett.* **22,** 79–84.

Moore, J. C., and Arnold, F. H. (1996). *Nature Biotechnology.* **14,** 458–467.

Moore, J. C., Jin, H.-M., Kuchner, O., and Arnold, F. H. (1997). *J. Mol. Biol.* **272,** 336–347.

Mühlenbein, H. (1992). *Parallel Problem Solving from Nature 2, Brussels.* 15–25.

Müller, H. J. (1964). *Mutat. Res.* **1,** 2.

Newman, M. E. J., and Engelhardt, R. (1998). *Proc. R. Soc. Lond. B.* **265,** 1333–1338.

Nowak, M., and Shuster, P. (1989). *J. Theor. Biol.* **137,** 375–395.

Ohta, T. (1973). *Nature.* **246,** 96–98.

Ohta, T. (1997). *J. Mol. Evol.* **44,** S9–S14.

Otto, S. P., Feldman, M. W., and Christiansen, F. B. (1994). In ''Frontiers in Mathmatical Biology'' (S. A. Levin, ed.), pp. 198–211, Springer-Verlag, Berlin.

Pàl, K. F. (1995). *Biol. Cybern.* **73,** 335–341.

Pawlowski, K., Bierzynski, A., and Godzik, A. (1996). *J. Mol. Biol.* **258,** 349–366.

Perelson, A. S. (1990). ''1989 Lectures in Complex Systems'' (E. Jen, ed.), Vol. II., pp. 465–499, Addison-Westley.

Perelson, A. S., and Macken, C. A. (1995). *Proc. Natl. Acad. Sci. USA.* **92,** 9657–9661.

Perelson, A. S., and Oster, G. F. (1979). *J. Theor. Biol.* **81,** 645.

Pollack, S. J., Jacobs, J. W., and Schultz, P. G. (1986). *Science.* **234,** 1570–1573.

Ponder, J. W., and Richards, F. M. (1987). *J. Mol. Biol.* **193,** 775–791.

Prügel-Bennett, A., and Shapiro, J. L. (1994). *Physical Review Letters.* **72,** 1305–1309.

Prügel-Bennett, A., and Shapiro, J. L. (1997). *Physica D.* **104,** 75–114.

Rattray, M., and Shapiro, J. (1997). In ''Foundations of Genetic Algorithms 4'' (R. K. Belew and M. D. Vose, eds.), pp. 117–139, Morgan Kaufmann, San Francisco.

Reeves, C. R. (1993). *Proceedings of the Fifth International Conference on Genetic Algorithms,* Urbana-Champaign. 92–99.

Reidhaar-Olson, J. F., and Sauer, R. T. (1988). *Science.* **241,** 53–57.

Reidhaar-Olson, J. F., and Sauer, R. T. (1990). *Proteins.* **7,** 306–316.

Reidys, C., Stadler, P. F., and Schuster, P. (1997). *Bull. Math. Biol.* **59,** 339–397.

Roberts, R. W., and Ja, W. J. (1999). *Curr. Opin. Struct. Biol.* **9,** 521–529.

Saven, J. G., and Wolynes, P. G. (1997). *J. Phys. Chem. B.* **101,** 8375–8389.

Schachman, H. K. (1993). *Curr. Opin. Struct. Biol.* **3,** 960–967.

Schaffer, J. D., Eshelman, L. J., and Offutt, D. (1991). In ''Foundations of Genetic Algorithms'' (G. J. E. Rawlings, ed.), pp. 102–112, Morgan Kaufmann, San Mateo.

Schaffer, J. D., and Morishima, A. (1987). *Proceedings of the Second International Conference on Genetic Algorithms, Cambridge.* 36–40.

Schober, A., Thuerk, M., and Eigen, M. (1993). *Biol. Cybern.* **69,** 493–501.

Schuster, P., Fontana, W., Stadler, P. F., and Hofacker, I. L. (1994). *Proc. Roy. Soc.* (*London*) *B.* **255,** 279–284.

Schuster, P., and Stadler, P. F. (1994). *Computers Chem.* **18,** 295–324.

Schuster, P., and Stadler, P. F. (1996). *SFI Proceedings for the HIV & HPV Structures Workshop,* Santa Fe, New Mexico. 1–30.

Shakhnovich, E. I. (1994). *Phys. Rev. Lett.* **72,** 3907–3910.

Shao, X., Zhao, H., Giver, L., and Arnold, F. H. (1998). *Nucleic Acids Res.* **26,** 681–683.

Shao, Z., and Arnold, F. H. (1996). *Curr. Opin. Struct. Biol.* **6,** 513–518.

Sherrington, D., and Kirkpatrick, S. (1975). *Phys. Rev. Lett.* **35,** 1792–1795.

Shoichet, B. K., Baase, W. A., Kuroki, R., and Matthews, B. W. (1995). *Proc. Natl. Acad. Sci. USA.* **92,** 452–456.

Skandalis, A., Encell, L. P., and Loeb, L. A. (1997). *Chemistry & Biology.* **4,** 889–898.

Skinner, M. M., and Terwilliger, T. C. (1996). *Proc. Natl. Acad. Sci. USA.* **93,** 10753–10757.

Spears, W. M. (1993). In ''Foundations of Genetic Algorithms 2'' (L. D. Whitley, ed.), pp. 221–237, Morgan Kauffman, San Mateo.

Spears, W. M., and De Jong, K. A. (1991). *Proceedings of the Fourth International Conference on Genetic Algorithms,* San Diego. 230–236.

Spiller, B., Gershenson, A., Arnold, F. H., and Stevens, R. C. (1999). *Proc. Natl. Acad. Sci. USA.* **96,** 12305–12310.

Stadler, P. F. (1992). *Europhys. Lett.* **20,** 479–482.

Stadler, P. F. (1996). *J. Math. Chem.* **20,** 1–45.

Stadler, P. F., and Grüner, W. (1993). *J. Theor. Biol.* **165,** 373–388.

Stemmer, W. P. C. (1994a). *Proc. Natl. Acad. Sci. USA.* **91,** 10747–10751.

Stemmer, W. P. C. (1994b). *Nature.* **370,** 389–391.

Strait, B. J., and Dewey, T. G. (1996). *Biophys. J.* **71,** 148–155.

Sumida, B. H., Houston, A. I., McNamara, J. M., and Hamilton, W. D. (1990). *J. Theor. Biol.* **147,** 59–84.

Sun, F. (1999). *J. Comp. Biol.* **6,** 77–90.

Suzuki, M., Christians, F. C., Kim, B., Skandalis, A., Black, M. E., and Loeb, L. A. (1996a). *Molecular Diversity.* **2,** 111–118.

Suzuki, M., Baskin, D., Hood, L., and Loeb, L. A. (1996b). *Proc. Natl. Acad. Sci. USA.* **93,** 9670–9675.

Syswerda, G. (1989). *Proceedings of the Third International Conference on Genetic Algorithms,* George Mason University. 2–9.

Tachida, H. (1991). *Genetics.* **128,** 183–193.

Tianna, G., Broglia, R. A., Roman, H. E., Vigezzi, E., and Shakhnovich, E. (1998). *J. Chem. Phys.* **108,** 757–761.

Trakulnaleamsai, S., Yomo, T., Yoshikawa, M., Aihara, S., and Urabe, I. (1995). *J. Ferment. Bioeng.* **79,** 107–118.

van Nimwegen, E., and Crutchfield, J. P. (1999a). *Machine Learning.* SFI Preprint 98-10-090.

van Nimwegen, E., and Crutchfield, J. P. (1999b). *Bull. Math. Biol.* SFI Preprint 99-07-041.

van Nimwegen, E., Crutchfield, J. P., and Huynen, M. (1999). *Proc. Natl. Acad. Sci. USA.* **96,** 9716–9720.

Voigt, C. A., Arnold, F. H., and Wang, Z.-G. (2000a). To be published.

Voigt, C. A., Mayo, S. L., Arnold, F. H., and Wang, Z.-G. (2000b). *Proc. Natl. Acad. Sci. USA,* submitted.

Vose, M. D., and Liepens, G. E. (1991). *Proceedings of the Fourth International Conference on Genetic Algorithms,* San Diego. 237–242.

Wagner, A. (1996). *Evolution.* **50,** 1008–1023.

Wagner, A., and Stadler, P. F. (1999). *J. Exp. Zool.* **285,** 119–133.

Wagner, G. P., and Gabriel, W. (1990). *Evolution.* **44,** 715–731.

Wang, J., Onuchic, J., and Wolynes, P. (1996). *Phys. Rev. Lett.* **76,** 4861–4864.

Wang, Q. (1987). *Biol. Cybern.* **57,** 95–101.

Weinberger, E. (1990). *Biol. Cybern.* **63,** 325–336.

Weinberger, E. (1991). *Phys. Rev. A.* **44,** 6399–6413.

Wells, J. A. (1990). *Biochemistry.* **29,** 8509–8517.

Wilde, J. A., Bolton, P. H., Dell'Acqua, M., Hilber, D. W., Pourmotabbed, T., and Gerlt, J. A. (1988). *Biochemistry.* **27,** 4127–4132.

Wilke, C. O., and Martinetz, T. (1999). *Phys. Rev. E.* **50,** 2154–2159.

Wright, S. (1932). *Proceedings of the Sixth International Congress on Genetics.* 356–366.

Wright, S. (1967). *Proc. Natl. Acad. Sci. USA.* **58,** 165–172.

Xu, J., Deng, Q., Chen, J., Houk, K. N., Bartek, J., Hilvert D., and Wilson, I. A. (1999). *Science.* **286,** 2345–2348.

You, L., Arnold, F. H. (1994). *Protein Engineering* **9,** 77–83.

Youvan, D. C., Arkin, A. P., and Yang, M. M. (1992). In ''Parallel Problem Solving from Nature 2'' (R. Männer and B. Manderick, eds.), pp. 401–410, Elsevier Science Publishers.

Zhang, X., Baase, W. A., Shoichet, B. K., Wilson, K. P., and Matthews, B. W. (1995). *Protein Engineering.* **8,** 1017–1022.

Zhang, Y.-C. (1997). *Phys. Rev. E.* **55,** R3817–R3819.

Zhao, H., and Arnold, F. H. (1997a). *Curr. Opin. Struct. Biol.* **7,** 480–485.

Zhao, H., and Arnold, F. H. (1997b). *Nucleic Acids Res.* **25,** 1307–1308.

Zhao, H., and Arnold, F. H. (1999). *Protein Engineering.* **12,** 47–53.

Zhao, H., Giver, L., Shao, Z., Affholter, J. A., and Arnold, F. H. (1998). *Nat. Biotechnol.* **16,** 258–261.

TEMPERATURE ADAPTATION OF ENZYMES: LESSONS FROM LABORATORY EVOLUTION

By PATRICK L. WINTRODE and FRANCES H. ARNOLD

Division of Chemistry and Chemical Engineering 210-41, California Institute of Technology, Pasadena, California 91125

I. Introduction

Life on earth flourishes over a wide range of temperatures. While any one species usually thrives in a much narrower range, an extraordinary diversity of species spans the biological temperature scale, from ~ −10 to +115°C. Life's biochemical building blocks, and in particular its protein catalysts, also function optimally within rather narrow temperature ranges. Adaptation to different temperature niches is therefore dependent on adaptation of the molecular components (Hochachka and Somero, 1984). This provides a marvelously rich source of protein engineering data for the biochemist bent on uncovering the secrets of protein function.

Most biochemical studies have focused on enzymes from mesophiles, organisms adapted for life at moderate temperatures (~20°–40°C)

0065-3233/00 $35.00

(Fig. 1). However, organisms living at ''extreme'' temperatures, including thermophiles—primarily bacteria and archaea (~60°–80°C)—extreme thermophiles, or hyperthermophiles (>80°C), and psychrophiles—bacteria, plants, fungi, and certain fish species that live at temperatures below ~15°C and often as low as 0°C—have fascinating stories to tell. Driving the growing interest in studying extremophiles is the dual hope of discovering new enzymes for biotechnology as well as new insights into enzyme function. Enzymes from thermophilic organisms are stable and catalytically active at temperatures at which meso-

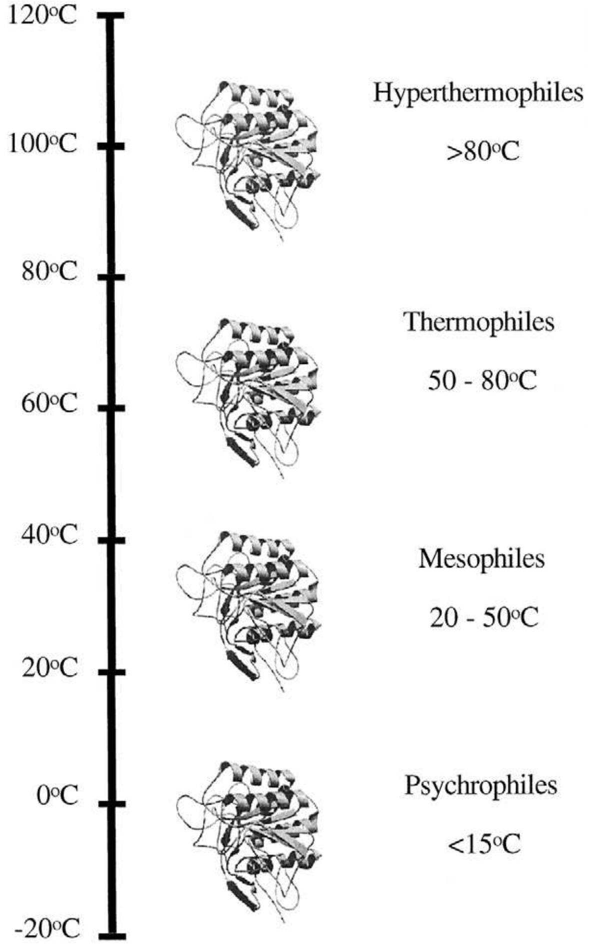


FIG. 1. Optimal growth temperatures for mesophilic and extremophilic organisms.

philic enzymes rapidly denature. Enzymes from psychrophilic organisms catalyze reactions at low temperatures at rates that are comparable to those of mesophilic enzymes at their physiological temperatures, despite the intrinsically slower rates of chemical reactions at low temperature. The mere existence of an enzyme that functions at 100°C demands we ask where this remarkable stability comes from. At the other end of the spectrum, we wonder how enzymes from psychrophiles can be so catalytically efficient at the freezing point of water.

Numerous studies have characterized and compared the sequences, physiochemical properties, and structures of extremophilic enzymes with those of related mesophiles. Although exciting and often fruitful, this approach is also fraught with danger. The reason is best summarized by Dobzhansky's famous assertion: ''Nothing in biology can be understood except in the light of evolution'' (Dobzhansky, 1973). Proteins are shaped by their histories (not unlike people). And knowing these evolutionary histories can help us to understand their behavior. But evolution is the source of much confusion to the biochemist who is looking for a reason for this or that trait. Sometimes there simply is no reason (a trait is an historical accident), and sometimes we cannot know the reason because we do not know what past selective pressures have influenced the properties of an enzyme being studied in the present.

As the products of long and complicated evolutionary trajectories, the bulk of which we will never know, the molecules of nature will always present confusing and contradictory testimony. Evolution may come to the rescue, however. Evolutionary protein design methods—mimicking Darwinian evolution in the test tube—may help remove some of the confusion. First, the set of molecules *not* found in nature make up an even richer bank of molecular diversity and function than the natural ones. Exploring this set allows us to distinguish the physically possible from the biologically relevant, which is a key step in removing the confusion of evolution. Studies of natural enzymes from mesophilic and extremophilic sources establish which protein properties (and combinations of properties) are biologically acceptable at different temperatures. Laboratory evolution frees the enzyme from constraints imposed by the need to function in a living organism and expands the inquiry to the *physically* possible. Second, it may be possible to mimic evolutionary pressures at work in nature, in order to recreate the processes by which natural molecules have come about. The laboratory offers well-defined conditions for the evolution of specific traits, and the intermediate products are there to analyze. To the extent that these experiments have relevance to biological evolution, we may be able to retrace history. Finally, with laboratory evolution, the conditions are controlled, the selective pressures well defined, and products can acquire significant functional changes with minimal changes in sequence, greatly simplifying the process of relating observed differences in behavior to specific differences in structure or sequence.

In this chapter, we will outline how evolutionary protein design methods are now being used to help uncover the molecular basis for temperature adaptation in enzymes. Before doing this, however, we will briefly review how temperature affects protein stability and enzyme activity. Then we will discuss some of the results of comparative studies of enzymes isolated from organisms adapted to different temperatures and the questions that can be addressed by laboratory evolution.

II. Influence of Temperature on Enzymes

A. *Stability*

Stability refers to the ability of a protein to maintain its native three-dimensional structure in the face of thermal fluctuations, chemical denaturants, or other factors. The term ''native state'' does not refer to a single unique conformation. Although the data from structural studies are generally presented as a single set of three-dimensional atomic coordinates, the native state is in fact an ensemble of related conformations. Experimentally determined structures represent the average conformation. The denatured state also consists of many conformations, although, unlike the native state, there is no single well-defined average structure. It has been observed that the transition from the native state to the denatured state is highly cooperative, resembling a first order phase transition (Privalov, 1979).

Because the native and denatured states are both well-defined thermodynamically, protein stability can be expressed as the difference in Gibbs free energy (ΔG_{unfold}) between the two. The equilibrium between the native and denatured states is temperature dependent. Plotted as a function of temperature, ΔG_{unfold} follows a parabolic curve, crossing zero at both low and high temperatures with a maximum somewhere in between (Privalov, 1979). The points at which $\Delta G_{\text{unfold}} = 0$ define the denaturation temperatures for a protein under a given set of conditions. The cold denaturation temperature is generally well below the freezing point of water for globular proteins and is thus of limited relevance to most biological situations (it may be of greater importance, however, for the stability of multimeric proteins—see Jaenicke, 1990). More important in most cases is the high temperature denaturation point. Because $\Delta G = \Delta H - T\Delta S$, the shape of the free energy curve for proteins is determined by the temperature dependence of the thermodynamic parameters ΔH_{unfold} and ΔS_{unfold}. The relative contributions of changes in the enthalpy and the entropy of unfolding vary both quantitatively and qualitatively with temperature. At high temperatures, ΔH_{unfold} and ΔS_{unfold} are both positive. Proteins at high temperatures are thus stabilized by enthalpic interactions. As temperature is decreased, the entropy of denaturation becomes negative, and $T\Delta S_{\text{unfold}}$ becomes a stabilizing factor. As temperature is further decreased, ΔH_{unfold} becomes negative and the protein is enthalpically destabilized and entropically stabilized, the opposite of the situation at high temperatures (Privalov, 1979).

Excluding covalent linkages, the native structures of proteins are maintained by large numbers of weak noncovalent interactions, which include van der Waals contacts, hydrogen bonds between buried groups, and salt bridges, as well as protein-solvent interactions arising from the burial or exposure of polar and nonpolar groups. The relative contributions these forces make to protein stability are still not fully understood, and because they all involve entropic and enthalpic components, their temperature dependence is complex. It has long been known that the removal of nonpolar groups from water and their burial in the protein interior is a critical factor in stabilizing folded proteins (Kauzmann, 1959). Previously it was believed that this was due almost entirely to the solvation properties of nonpolar groups: the dissolution of nonpolar compounds in water results in a significant decrease in entropy and is thus unfavorable. More recently, it has been shown that a large fraction of the stabilization resulting from the burial of nonpolar groups comes from the formation of enthalpically favorable van der Waals contacts between these groups in the protein interior (Makhatadze and Privalov, 1995). The other major contributor to the stability of the native state is thought to be the formation of hydrogen bonds between buried polar groups (Makhatadze and Privalov, 1995; Pace *et al.,* 1996). Buried salt bridges are thought to contribute little to stability (Hendsch and Tidor, 1994; Waldburger *et al.,* 1995). There is some evidence, however, suggesting that buried salt bridges may be stabilizing at higher temperatures due to changes in the solvation free energies of charged groups (Elcock, 1998). Studies on model compounds indicate that the free energy of solvation of nonpolar groups is unfavorable and becomes more unfavorable with increasing temperature, while the solvation of polar groups is favorable, becoming less favorable with increasing temperature (Privalov and Gill, 1988; Makhatadze and Privalov, 1995).

The total free energy of stabilization for proteins is, even at the temperature of maximum stability, quite small. For the majority of proteins studied so far, the ΔG_{unfold} at 25°C falls between 8 and 17 kcal/mol. This free energy is equivalent to only a few hydrogen bonds out of the hundreds that exist in a typical folded protein. From this it follows that very minor adjustments in the number and type of interactions in a protein can result in significant changes in stability. Indeed, the potential of subtle changes to significantly alter stability is partially responsible for the difficulty of identifying stabilization strategies in thermophilic proteins.

The covalent structure of the protein can also be affected at high temperatures, leading to deactivation and degradation. At temperatures approaching 100°C, proteins and peptides undergo a number of irreversible degradation reactions. These include the deamidation of the amide side chains of Asn and Gln, succinimide formation at Glu Asp, and oxidation of His, Met, Cys, Trp and Tyr residues (Daniel and Danson, *Methods in Enzymology,* in press). There is evidence (Hensel *et al.,* 1992) that the folded conformation provides substantial protection against these degradation mechanisms even at high temperatures, suggesting that degradation reactions may be preceded by unfolding (Daniel and Danson, *Methods in Enzymology,* in press).

B. Catalytic Activity

The Arrhenius equation describes temperature dependence of the rate constant of a chemical reaction:

$$k = A\mathrm{e}^{-\Delta E^*/RT}$$

where A is a preexponential factor, R is the gas constant, T is the absolute temperature, and ΔE^* is the activation energy. Thus, an exponential increase in the reaction rate is expected with increasing temperature, until competing processes (e.g., boiling of the solvent, breakdown of reactants or products) become significant. The rates of reactions catalyzed by enzymes are also temperature dependent, but the behavior is more complex. Typically, the rate of an enzymatic reaction will increase two or threefold with each 10°C temperature increase. This increase will continue until the onset of denaturation, at which point the reaction rate decreases. Denaturation leads to an apparent temperature optimum (T_{opt}) for an enzyme-catalyzed reaction. The value of T_{opt} depends on how it is measured, in particular on how long the enzyme is incubated at the elevated temperature. Increasing the incubation time often leads to a decrease in T_{opt} as a result of irreversible denaturation. T_{opt} is therefore of limited biological relevance, although it does generally show a correlation with physiological temperature. Enzymes from thermophilic organisms tend to have higher temperature optima than their mesophilic or psychrophilic homologs.

Substrate binding is also affected by temperature. Protein-substrate interactions are mediated by the same weak forces that determine protein stability, and as the magnitudes of these forces change with temperature so does the strength of substrate binding. Studies have found that, while the enthalpy and entropy of substrate binding can depend quite strongly on temperature, they often change in a compensatory manner, leading to a much smaller temperature dependence for binding free energy (Hinz and Jaenicke, 1975; Schmid *et al.,* 1976). Binding can also be influenced by the fact that diffusion slows as the temperature is lowered. This may be offset by the fact that once the enzyme-substrate complex is formed, it is less likely to be disrupted by thermal fluctuations at lower temperatures.

III. Studies of Natural Extremophilic Enzymes

Proteins from extremophilic organisms, particularly thermophiles, have been the subject of intensive research in recent years. This work has been the subject of numerous reviews (Jaenicke and Bohm, 1998; Russel and Taylor, 1995; Vogt and Argos, 1997; Gerday *et al.,* 1997; Somero, 1995), and we will make no attempt at an in-depth summary. We will confine ourselves to briefly stating the major trends identified thus far. Explaining these trends becomes complicated because the many weak interactions that determine enzyme stability and activity have complex temperature dependencies (see Section II). And evolution injects considerable confusion beyond the laws of physical chemistry.

A. *Sequence and Structure*

Attempts to identify the factors responsible for the remarkable stability of thermophilic enzymes have included sequence comparisons, studies focusing on the thermodynamics of denaturation, structural comparisons, and molecular dynamics simulations. Comparisons of the sequences and structures of thermophilic proteins with their mesophilic homologs indicate that nature relies on no single strategy for stabilization (Querol *et al.,* 1996; Vielle and Zeikus, 1996; Adams and Kelly, 1998; Jaenicke and Bohm, 1998). Relative to their mesophilic counterparts, thermophilic proteins show enhanced internal packing of nonpolar groups (Yip *et al.,* 1995; Yip *et al.,* 1998; Auerbach *et al.,* 1998), shortened surface loops (Russell *et al.,* 1998; Thompson and Eisenberg, 1999), stabilization of helices (Macedo-Ribeiro *et al.,* 1996; Henning *et al.,* 1995;

Zaiss *et al.,* 1998; Auerbach *et al.,* 1998; Auerbach *et al.,* 1997), greater numbers of hydrogen bonds (Macedo-Ribeiro *et al.,* 1996; Wallon *et al.,* 1997; Aoshima and Oshima, 1997) and larger networks of ion pairs (Russell *et al.,* 1997; Hatanaka *et al.,* 1997; Yip *et al.,* 1995; Yip *et al.,* 1998; Korndorfer *et al.,* 1995; Hennig *et al.,* 1995; Akanuma *et al.,* 1998; Aoshima and Oshima, 1997; Auerbach *et al.,* 1998; Zaiss *et al.,* 1998; Auerbach *et al.,* 1997a; Auerbach, *et al.,* 1997b). The presence of larger ion pair networks, particularly in the interface regions of dimeric or multimeric proteins, is the most consistent feature of the thermophilic proteins studied thus far, but there does not appear to be any universally preferred stabilizing mechanism. These differences may merely reflect chance—some types of favorable interactions just happened to occur rather than others—or they may indicate that some stabilization strategies are favored by particular classes of protein folds.

Although less effort has been directed to the enzymes from psychrophiles, some well-studied cases exist. As expected, psychrophilic enzymes are less stable at moderate temperatures than their mesophilic counterparts. Differences between psychrophilic and mesophilic homologs include fewer salt bridges (Davail *et al.,* 1994; Narinx *et al.,* 1997; Aghajari *et al.,* 1998; Kim *et al.,* 1999) and aromatic-aromatic interactions (Davail *et al.,* 1994; Narinx *et al.,* 1997), fewer prolines in loop regions (Davail, *et al.,* 1994; Narinx *et al.,* 1997; Russell *et al.,* 1997; Aghajari *et al.,* 1998; Kim *et al.,* 1999), more hydrophilic surface residues (Davail *et al.,* 1994; Narinx *et al.,* 1997; Aghajari *et al.,* 1998; Kim *et al.,* 1999), a lower Arg/(Arg + Lys) ratio (Davail *et al.,* 1994; Narinx *et al.,* 1997; Aghajari *et al.,* 1998), and weaker binding of ligands, including metals and other ions (Davail *et al.,* 1994; Narinx *et al.,* 1997; Aghajari *et al.,* 1998; Kim *et al.,* 1999). In general, these differences indicate fewer stabilizing interactions in the cold-adapted enzymes. The lower Arg/Lys ratio, for instance, may be related to the fact that arginine is capable of forming more than one ion pair, due to the charge resonance of the guanidinium group (Feller and Gerday, 1997). Due to constraints on their allowed ϕ–ψ angles, prolines in loops tend to restrict mobility. Although higher mobility might be entropically stabilizing, large loop displacements have also been identified as initial events in protein unfolding (Caflisch and Karplus, 1995; Lazaridis *et al.,* 1997). Additionally, prolines reduce the available conformations and thus the entropy of the denatured state. Since stabilization is the difference in free energy between the native and unfolded states, a reduction in the entropy of the unfolded state is thermodynamically stabilizing. Binding of metal ions such as calcium dramatically stabilizes a number of proteins; weaker affinity for these ligands can decrease stability.

Comparisons of psychrophilic, mesophilic, and thermophilic enzymes suggest that a continuum of adjustments accompany adaptation to different temperatures (Davail *et al.,* 1994; Feller and Gerday, 1997). Relative to mesophiles, the same kinds of weakly stabilizing interactions that are found in greater proportion in thermophilic enzymes appear in fewer numbers in their psychrophilic counterparts.

B. Thermodynamics, Dynamics, and Function: "Corresponding States"

Due primarily to problems associated with reversibility, detailed thermodynamic studies on thermophilic proteins have been lacking. Several recent studies, however, have found that thermophilic proteins use various thermodynamic strategies for stabilization (Fig. 2). Ferredoxin from the thermophile *Thermotoga maritima* shows an unusually small heat capacity change (ΔC_p) on denaturation relative to mesophilic ferredoxins (Pfeil *et al.,* 1997). ΔC_p determines the temperature dependence of the thermodynamic parameters ΔH_{unfold} and ΔS_{unfold}, which in turn determine ΔG_{unfold}, and a decrease in ΔC_p broadens the free energy curve. This moves the upper denaturation point ($\Delta G_{\text{unfold}} = 0$) to higher temperature. In contrast, a study of dihydrofolate reductase from the same organism found no broadening of the free energy curve; rather the free energy of stabilization was increased overall, and the temperature of maximum

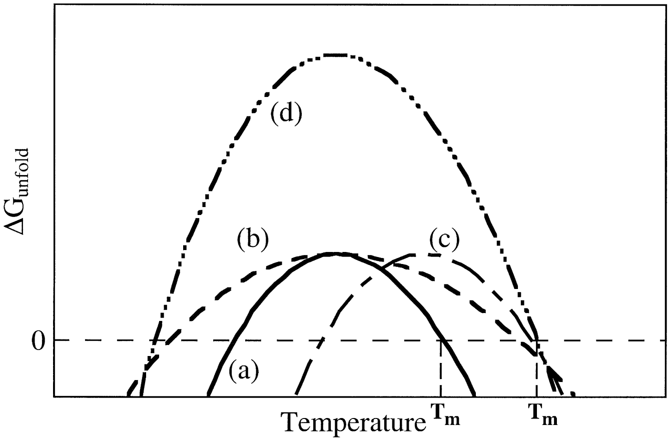


FIG. 2. Different thermodynamic strategies for increasing the denaturation temperature of a protein. (a) Free energy curve for a mesophilic enzyme. (b) Thermostabilization by broadening the free energy curve. (c) Thermostabilization by shifting the curve to higher temperatures. (d) Thermostabilization by increasing the free energy of stabilization at all temperatures.

stability was shifted to higher temperatures (Dams and Jaenicke, 1999). A comparison of cellulase domains from the thermophile *Thermomonospora fusca* and the mesophile *Cellulomonas fimi* also uncovered an increase in the maximum free energy of stabilization (Beadle *et al.,* 1999). Studies of a psychrophilic α-amylase found that the free energy profile was not shifted to lower temperatures, but that the molar free energy per residue (used for comparing the free energies of proteins of different sizes) was three to four times lower than those of mesophilic α-amylases over the entire temperature range in which it was stable (Feller and Gerday, 1999). This diversity of thermodynamic adaptations likely reflects the diversity of structural adaptations previously discussed.

Enzymes adapted to different temperatures often exhibit behavior that has been referred to as ''corresponding states'' (Jaenicke and Bohm, 1998). That is, equivalent enzymes from psychrophilic, mesophilic, and thermophilic organisms show comparable levels of activity at their respective physiological temperatures. At the same time, enzymes from thermophiles tend to be poorly active at moderate temperatures relative to their mesophilic counterparts (similar results are found for mesophiles relative to psychrophiles). This has also been found to be true for other properties. Several studies have shown that at their respective physiological temperatures the thermodynamic stability of thermophilic and psychrophilic enzymes, as measured by ΔG_{unfold}, is comparable to that of mesophilic enzymes at their physiological temperatures (Dams and Jaenicke, 1999). Substrate binding, as measured by K_M, is yet another attribute that has been found to be conserved at the physiological temperatures of enzymes adapted to different environments (Somero, 1995).

Measures of protein dynamics and mobility have also shown a similar trend. Varley and Pain (1991) measured the conformational flexibility of thermophilic and mesophilic 3-phosphoglycerate kinases by monitoring acrylamide-quenching of the fluorescence of a buried tryptophan residue. At 25°C, the quenching rate constant is smaller for the thermophilic PGK, indicating less penetration of acrylamide molecules into the core of the protein. Extrapolation to the thermophile's physiological temperature, however, indicated a quenching rate constant similar to that of the mesophilic PGK at 25°C. More recently, the conformational flexibilities of thermophilic and mesophilic 3-isopropylmalate dehydrogenases were estimated by hydrogen/deuterium exchange (Zavodzsky *et al.,* 1998). Similar to the results found for PGK, the thermophile showed less exchange than the mesophile at 25°C, but the two enzymes had similar exchange patterns at their respective temperature optima (70°C and 48°C). Computer simulations provide another means of examining the dynamic behavior of proteins. Molecular dynamics simulations of a mesophilic and hyperthermophilic rubredoxin (Lazaridis *et al.,* 1997) show that the per residue RMS fluctuations about the average structure were slightly smaller for the thermophile than for the mesophile near room temperature.

Observations such as these from related enzymes adapted to different temperatures have led to the assertion that there exists an intrinsic trade-off between stability at elevated temperatures and catalytic activity at lower temperatures, such that one property must necessarily come at the cost of the other. This alleged activity-stability trade-off is typically justified by reference to the opposing reliance of these properties on the conformational mobility, or flexibility, of the enzyme. Although the classical ''lock and key'' model of enzyme catalysis implied a rigid enzyme structure, it has been appreciated for many years that molecular motions play an essential role in substrate specificity and catalysis. At temperatures near ~200°K, myoglobin and ribonuclease A have been shown to undergo a ''dynamic transition,'' below which the large-scale collective motions are frozen out, and only harmonic motions of individual atoms remain (Doster *et al.,* 1989; Tilton *et al.,* 1992). Petsko and co-workers have shown that, below this transition point, ribonuclease A also loses its ability to bind substrates and inhibitors, thus demonstrating that collective protein motions are required for function (Rasmussen *et al.,* 1992). Although fluctuations in structure are essential for catalytic function, fluctuations that are too large will lead to disruption of the native structure and loss of activity. The magnitude of the fluctuations depends on the thermal energy, *kT,* available to the protein and thus increases with temperature.

The explanation most commonly offered, therefore, for the trade-off between stability and activity is that, during the course of evolution, enzymes have adjusted the strength and number of their stabilizing interactions so as to optimize the balance between rigidity (for stability) and flexibility (for activity) at their physiologically relevant temperatures (Varley and Pain, 1991; Davail *et al.,* 1994; Zavodszki *et al.,* 1998; Fontana *et al.,* 1998; Feller *et al.,* 1999; Hollien and Marqusee, 1999). Hence, to achieve catalytic efficiency comparable to that of a mesophilic enzyme at its natural physiological temperature, a psychrophilic enzyme functioning at 0°C must exhibit thermal motions of the same magnitude as those of the mesophilic enzyme at 37°C. When the psychrophilic enzyme is exposed to mesophilic temperatures, however, these thermal motions become so great that they can lead to the loss of native structure. When a mesophilic enzyme is cooled from 37°C to 0°C, the reduction in thermal fluctuations will so diminish the conformational mobility of the enzyme that its catalytic efficiency is compromised.

C. *The Confounding Effects of Evolution*

The argument previously outlined provides an appealing physiochemical explanation for the stability and activity behavior of homologous enzymes adapted to different temperatures. However, one cannot interpret the behavior of a biological system solely in physiochemical terms. All these enzymes are the products of evolution. While they are certainly subject to the laws of physics and chemistry, the evolutionary process imposes its own, additional constraints. We will see that the stability-activity trade-off is not a necessary characteristic of enzymes, especially not those evolved in the laboratory.

Comparative studies of mesophilic, thermophilic, and psychrophilic enzymes point to many interesting differences in sequence, structure, function, dynamics, and thermodynamic properties. It is by no means clear, however, which of these differences are central to the process of thermal adaptation and which are mere side effects—the results of neutral drift or even adaptation to other selective pressures. This points to a serious difficulty facing comparative studies: identifying which enzyme properties have evolved under selective pressure. If mutations arise as an adaptive response to one set of demands, say to the need to have higher activity at low temperature, then it would be erroneous to interpret them as responsive to another demand (even though they also generate that property). ''Design rules'' based on such interpretations would be seriously flawed. In addition, natural selection can make complex demands on enzyme properties. Extremophiles from deep sea hydrothermal vents, for example, must adapt to both high temperature *and* high pressure, while bacteria and archaea found in hot springs will be adapted to both high temperature and high acidity. Determining which sequence and structural differences represent adaptation to which selective pressure is a daunting task.

Another major obstacle to comparative studies of naturally occurring proteins is that most mutations are not responsive to adaptive pressures. According to the neutral theory of evolution (Kimura, 1968; Kimura, 1983), many, if not most, of the mutations seen at the molecular level are neutral—they do not give rise to significant changes in fitness (survival and reproduction of the organism) and are fixed in evolving populations by random processes. The net effect is that protein sequences are subject to a continuous accumulation of neutral mutations, or genetic drift, which increase the distance between sequences without necessarily increasing the distance between functions. A thermophilic enzyme typically shares 30% to 50% amino acid identity with its mesophilic homologs (Adams and Kelly, 1998), comparable to the identity shared by pairs of homologous mesophilic enzymes. At 40% identity, two 300-residue enzymes will differ at 180 positions, and most of these differences likely contribute little or nothing to temperature adaptation. Testing all the amino acid substitutions and their possible combinations to distinguish adaptive from neutral substitutions is clearly impossible.

Neutral mutations are neutral with respect to fitness. This does not mean they are neutral with respect to all enzyme behaviors. In fact, many 'neutral' mutations will be deleterious to stability, catalytic ability, or any other property that does not contribute directly to fitness. Properties not protected by the purifying effects of natural selection can change as mutations accumulate, but the process is random and contains little information that can be used to elucidate mechanisms (Benner and Ellington, 1990; Benner, 1989).

We have already noted that cold-active enzymes from psychrophilic organisms exhibit poor thermostability compared to their mesophilic counterparts. We are faced with three possible explanations for this confinement of natural enzymes to such a limited region of function space (Fig. 3). One is that high activity at low temperatures is physically incompatible with high thermostability, so that an enzyme must necessarily sacrifice one in order to achieve the other. In addition to this physiochemical explanation, however, we must consider two others that arise from evolutionary considerations. One is that highly stable enzymes are actually harmful to psychrophilic organisms, so that thermostability is actually subject to *negative* selective pressure. The other is that high thermostability is simply not required by organisms that live in cold environments, so that thermostability has decreased over time due to random drift. (If temperature adaptation happened in the opposite direction, from cold to hot, then high activity at low temperatures would no longer be required by thermophilic enzymes and that property would be lost.) Comparative studies of natural mesophiles and extremophiles are severely hampered by the inability to distinguish between these possible explanations.

IV. Directed Evolution

Techniques that allow researchers to ''evolve'' specific enzyme properties in the laboratory circumvent some of the obstacles we encounter when we study only the products of natural evolution. Natural evolution is limited by both the rate at which mutations occur and the rate at which they become fixed in the population, with the consequence that new properties can take many years to appear. Laboratory evolution (or ''directed evolution'') can focus on single proteins, and the time scale

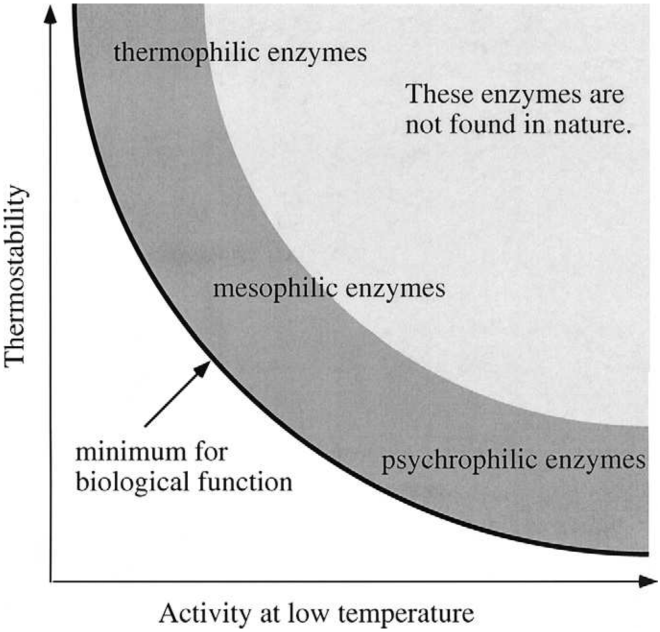


FIG. 3. Homologous enzymes adapted to different temperatures show a trade-off between catalytic activity at low temperatures (high for enzymes from psychrophilic organisms, but generally low for enzymes from thermophiles) and thermostability (high for thermophilic enzymes, but low for enzymes from psychrophiles). These natural enzymes lie in the darker shaded area, which is bounded on one side by the minimal stability and activity required for biological function. Enzymes that are both highly thermostable and highly active at low temperature (lighter shaded area) are generally not found in nature.

is greatly compressed by the experimenter, who takes control over all steps in the algorithm. Most important, and in contrast to natural evolution, the functional outcome is set by the experimenter, who controls the selection pressure(s). Using this control, we can investigate the acquisition of different traits, we can monitor the *process* by which this happens, and we can observe the functional consequences of acquiring specific traits. For example, we can probe the extent to which two properties (e.g., high temperature stability and low temperature activity) are coupled by evolving one property and observing the other. We can also attempt to evolve multiple properties simultaneously. We can explore properties and combinations of properties not observed in natural enzymes, in order to probe what is physically possible, free from the constraints of biologically relevant function. And, by analyzing the products of these experiments, we can hope to identify mechanisms for the acquisition of specific functional features. Since methods of directed evolution have been reviewed extensively (see, e.g., Gershenson and Arnold, 2000; Arnold and Wintrode, 1999), we will present only a brief overview before addressing issues relevant to thermal adaptation of enzymes.

A. *Evolutionary Strategy and Mutagenesis Methods*

Our preferred evolutionary strategy for probing the molecular basis of temperature adaptation is a simple Darwinian (adaptive) process: the accumulation over multiple generations of single, beneficial mutations produced randomly and fixed by selective pressure (Fig. 4). Accumulating single rather than multiple amino-acid substitutions per generation serves two important goals: it provides a near-optimal ''hit rate'' (fraction of clones exhibiting the desired property), and it removes all ambiguity in assigning functional mutations. The optimal hit rate depends on the fraction of amino-acid substitutions that do not influence the evolving property (neutral mutations). When this number is large, a higher mutation rate can be tolerated (Morawski *et al.,* 2000; see chapter by Voigt *et al.,* in this volume). However, when multiple mutations appear in a single generation, we cannot know which one is responsible for the change in phenotype without further experimentation. Backcrossing (Stemmer, 1994a) the evolved gene with wild type or another ancestral sequence can distinguish functional from neutral mutations (Zhao and Arnold, 1997a). Site-directed mutagenesis is not necessarily conclusive because some mutations may be beneficial in one background (e.g., the evolved sequence), but not in another (e.g., wild type) (Gershenson *et al.,* 2000; Spiller *et al.,* 1999).

Of course, this uphill random walk involving single steps is just one of the many ways in which evolution can proceed, in nature and in the laboratory. Our experience, however, is that this algorithm is highly efficient for improving functional traits such as catalytic activity or thermostability. Other approaches, including ''molecular breeding'' by recombination of homologous genes (see chapter by Ness *et al.* in this volume) and directing mutations (at higher mutagenesis rates) to specific regions of the gene (Kast and Hilvert, 1997; Skandalis *et al.,* 1997), can also generate enzymes with the desired functional traits. In fact, they may provide functional solutions that cannot be achieved with a single-step walk (Arnold, 1998). However, it is much more difficult to use the products of these evolutionary design processes to discover molecular mechanisms—the difficulty of assigning functional mutations returns us to the problems encountered in comparing natural, homologous enzymes.

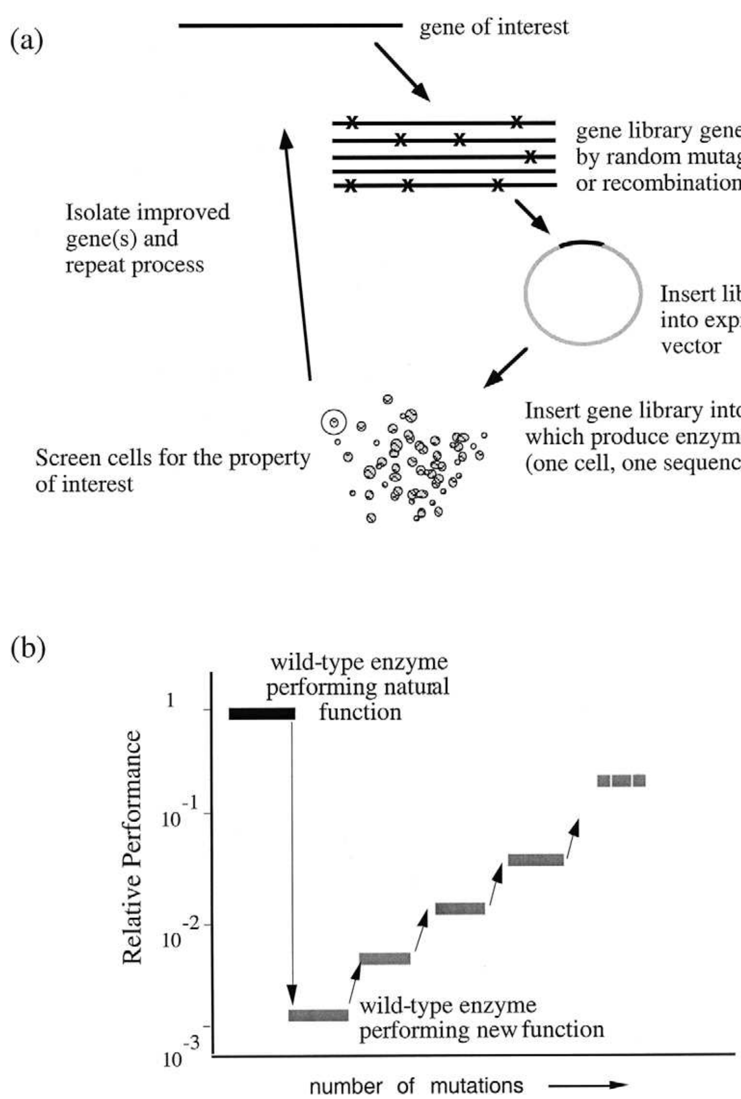

(a)
gene of interest
gene library generated by random mutagenesis or recombination
Insert library into expression vector
Insert gene library into host cells which produce enzyme variants (one cell, one sequence)
Screen cells for the property of interest
Isolate improved gene(s) and repeat process
(b)
wild-type enzyme performing natural function
wild-type enzyme performing new function
Relative Performance
1
10^{-1}
10^{-2}
10^{-3}
number of mutations

FIG. 4. (a) Experimental strategy for directed evolution. (b) Small improvements in fitness resulting from single amino-acid substitutions are accumulated over successive generations.

The random mutant library is typically generated using an error-prone polymerase chain reaction (PCR) in which the point mutation frequency is controlled by varying either the pH, the concentration of divalent cations, or other parameters (Leung *et al.,* 1989; Chen and Arnold, 1993; Caldwell and Joyce, 1994). A point mutation rate set to approximately three base mutations per gene generates the desired library of enzymes with primarily single amino-acid substitutions. Of course, a significant fraction of the library will have zero amino-acid substitutions, and some clones will have two or more. Random mutations can also be introduced by DNA shuffling (Stemmer, 1994a, 1994b) at a controllable rate (Zhao and Arnold, 1997b), as well as by other methods, including treatment with mutagens such as hydroxylamine or exposure to UV radiation. We prefer error-prone PCR because it is easy to use and mutation rate is easy to control.

A drawback to all random point mutation methods is the limited access to amino-acid diversity. At low mutagenesis rates, the probability that a single codon will experience multiple mutations is extremely small, with the consequence that only those amino-acid changes that can be accessed with a single base substitution can be explored in a single generation. Due to the structure of the genetic code, the amino-acid substitutions accessible by single mutations are relatively conservative. Aromatic amino acids are unlikely to replace aliphatic ones, for example, and large changes in polarity are rare. The inherent bias of PCR for transition mutations over transversions further limits the sequence space accessible in this evolutionary search. At present, the entire complement of twenty natural amino acids can be reached only by saturation mutagenesis focused on individual sites.

There is some evidence to support the notion that rapid adaptation of function is better achieved by having access to more nonconservative amino-acid substitutions. Still, point mutation methods such as error-prone PCR remain most suitable for attempts to mimic natural evolution or to retrace history. Further functional adaptation may be achieved by saturation mutagenesis at sites identified during random mutagenesis experiments (Miyazaki *et al.,* 1999).

The direction of evolution is dictated by the experimenter in the form of a screen (or selection) for the desired phenotype. Screening strategies are discussed below and in more detail by Gershenson and Arnold (2000). The single-step nature of the evolutionary walk, however, also places some limitations on the properties that can be evolved. Properties (or combinations of properties) that require multiple, simultaneous amino acid substitutions will not be acquired.

B. Screening Strategies for Temperature Adaptation

Screening the mutant libraries for improved variants is usually the most time-consuming part of the directed evolution experiment. It is also the most important: details of the screen will strongly influence the functional outcome. The goals of the experiment will dictate what temperature-dependent properties, thermostability, catalytic activity, etc., will be targeted in the screen and how changes will be monitored. We will discuss some important general considerations for designing screens for these properties. Specific examples of screening strategies used to probe different features of temperature adaptation and the results of those studies are presented in the following section. Because enzyme thermostability is conveniently measured using catalytic activity, thermostability and activity are often measured in the same screen.

Screening libraries of reversibly and irreversibly denaturing enzymes involves fundamentally different strategies. A convenient measure of thermostability for enzymes that undergo irreversible unfolding is retention of catalytic activity after incubation at high temperature. Screens based on residual activity are easily implemented in a standard 96-well plate format and have been used extensively in this laboratory for the directed evolution of thermostable enzymes (Giver *et al.,* 1998; Zhao and Arnold, 1997a, 1999; Miyazaki *et al.,* 2000). A typical procedure and results are shown in Figure 5. Single colonies are picked and grown in 96-well plates for enzyme expression. Cells (or supernatant, if the enzyme is secreted) are then transferred to two replica plates. One is assayed for activity (at a convenient assay temperature), while the other is incubated at high temperature for a fixed period of time and then assayed at the assay temperature. Assays are typically performed by monitoring a change in optical density, using a microtiter plate reader that can follow reaction kinetics in 96 or 384 wells simultaneously.

The ratio of the residual activity after incubation to the initial activity provides a measure of stability. Because it is a ratio of activities, this stability index automatically corrects for differences in activity due to variation in enzyme expression levels in the random mutant library. The majority of random mutants are less stable than the parent. However, a small number of more stable mutants can be identified provided the variation inherent in the screen is small (Gershenson and Arnold, 2000).

Although assays based on the retention of room temperature activity after heat treatment are convenient, the constraints imposed during the laboratory evolution can be quite different from anything an enzyme in a natural thermophilic organism might have encountered. Retaining activity at moderate temperatures is probably not relevant for enzymes

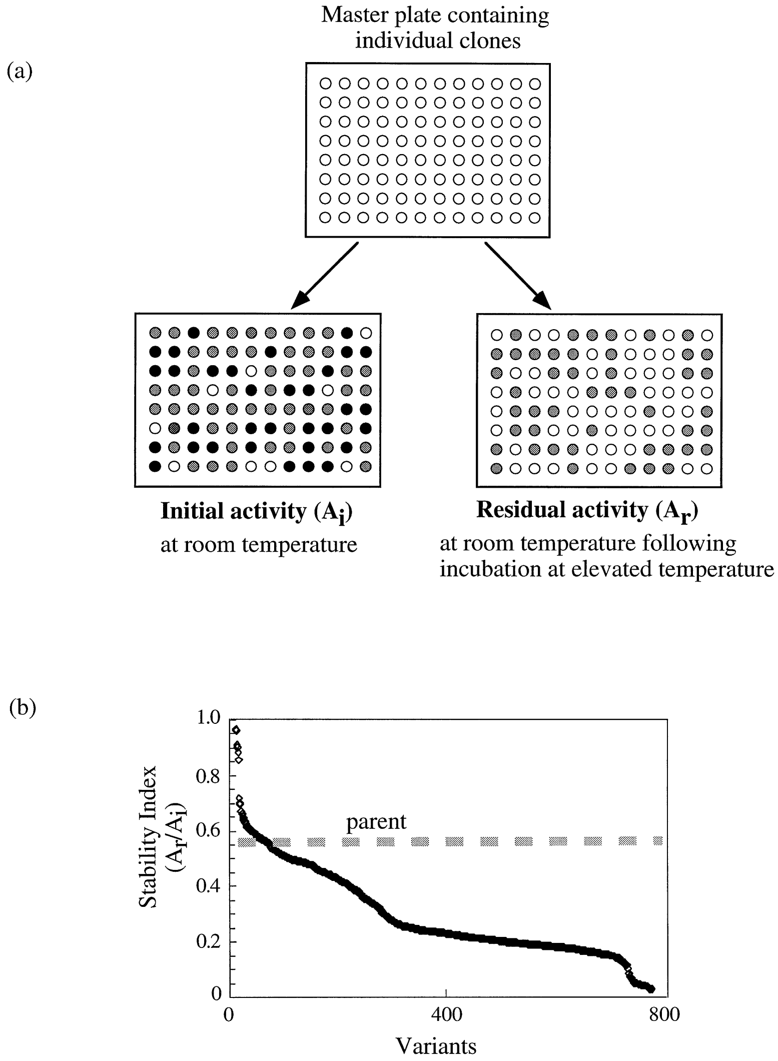


FIG. 5. (a) Experimental strategy for screening the thermostability of irreversibly denaturing enzymes. (b) Example of typical results for a mutant library; mutants are sorted according to stability.

in organisms that live at elevated temperatures. Likewise, activity at high temperature, which is presumably subject to direct selective pressure in thermophiles, is not measured in a residual activity assay performed at room temperature. The resulting thermostable mutants may therefore behave quite differently from thermophilic enzymes (see below). In addition, residual activity after heat treatment is a measure of the kinetics of irreversible inactivation rather than thermodynamic stability. Although the connection between these two measures of stability is not established, they are often closely correlated. For subtilisin BPN′, for example, resistance to irreversible inactivation and thermodynamic stability are correlated over a wide range of denaturing conditions (Pantoliano *et al.,* 1989).

Barring aggregation, enzymes that unfold reversibly will refold when cooled. Thus, their residual activity will not reflect thermostability. For these enzymes, catalytic activity must be measured at high temperature: a more thermostable enzyme retains catalytic activity at higher temperature, while loss of activity indicates unfolding. Most plate readers, however, do not go above ~50°C, precluding kinetic assays without specialized equipment. However, activity can be measured with an end-point assay, in which the reaction proceeds for a fixed time at high temperature and is then stopped by the addition of an inhibitor, a change in pH or solvent, or some other appropriate method. Total substrate depletion (or product formation) is measured. If an end-point assay is used, the reaction time and the substrate and enzyme concentrations must be chosen carefully so that the reaction is stopped while substrate depletion/product formation is increasing linearly with time. To correct for differences in protein expression levels, a replica plate can be assayed at room temperature, and the ratio of activities at the two temperatures can be used as a measure of stability.

Since most enzyme assays are done in aqueous solutions, measuring enzyme activity close to or above the boiling point of water presents special difficulties. First, the substrate must be stable at the elevated temperature so that degradation does not compete with conversion by the enzyme. A second consideration is the choice of assay buffer. The pK's of many commonly used biological buffers, such as TRIS, depend strongly on temperature, leading to complicating pH effects as the temperature is shifted appreciably from room temperature. A list of buffers suitable for use at high temperatures as well as an extensive discussion of other factors to consider when making high-temperature activity measurements can be found in Daniel and Danson (*Methods in Enzymology,* in press). Screening libraries for activity or resistance to denaturation above the boiling point also requires special high pressure plates.

If the enzyme of interest is essential for growth or survival, then it may be thermostabilized by selection in a thermophilic host (Oshima, 1994). The host growth temperature must be higher than the denaturation temperature of the enzyme to select for thermostable mutants. However, if the lower limit for growth is too far above the enzyme's denaturation temperature, the selection will require such large increases in stability that the desired mutants will be impossible to find. *Bacillus stearothermophilus,* which can grow from 55°C up to 70°C, and *Thermus thermophilus,* which can grow at 55°C to 80°–85°C, are useful host strains for the evolution of thermostability (Oshima, 1994). Strains lacking the gene coding for the enzyme of interest must be constructed. Mutations may be introduced by the *in vitro* techniques discussed above, and the mutagenized genes inserted into the host. Alternatively, the gene may be mutagenized *in vivo* through spontaneous errors in DNA replication. This latter method was used by Oshima and co-workers (Akanuma *et al.,* 1998) to generate thermostable 3-isopropylmalate dehydrogenases (discussed in Section IV, B, 4). Because evolution in a thermophilic host requires that the enzyme is functional in a biological context (i.e., contribute to the fitness of an organism), the selective pressures it experiences may be closer to those encountered during natural evolution. This can be an advantage or a disadvantage, depending on the goals of the experiment. For example, *in vivo* evolution is inappropriate for exploring nonnatural combinations of properties.

For some experiments, the goal may be to alter the activity of an enzyme at a given temperature, for example to increase it at low temperatures where mesophilic and thermophilic enzymes are generally poorly active. In principle, screening for low-temperature activity is straightforward: one simply looks for mutants showing increased activity at the temperature of interest. As with high-temperature measurements, however, there may be complications. It is not always possible to assay reaction kinetics at low temperatures, and an end-point assay may be required. More important when assaying for improved activity (at both low and high temperatures) is to select the appropriate screening substrate. The golden rule of directed evolution—"you get what you screen for"—dictates that if an enzyme is evolved for activity against a particular substrate, the final product of evolution will show improved activity toward that substrate. Activity toward other substrates may be increased, unchanged, or even diminished. The conditions of the screens for activity or thermostability can be altered as appropriate to probe other properties, including substrate or product inhibition, pH profiles, or specific inactivation mechanisms.

C. Directed Evolution of Thermostability

1. p-Nitrobenzyl Esterase

The thermostability of an α/β hydrolase from *Bacillus subtilis* was increased substantially by the accumulation of mutations over eight generations (Giver *et al.,* 1998; Gershenson *et al.,* 2000). The enzyme known as *p*-nitrobenzyl esterase (*p*NB E) is active toward ester intermediates in the synthesis of antibiotics that are protected with the *p*-nitrobenzyl group (Moore and Arnold, 1996). Eight rounds of mutagenesis/recombination and screening yielded *p*NB E variant 8G8 whose melting temperature of 69.5°C is a full 17°C above that of wild type.

The high thermostability of natural thermophilic enzymes is usually accompanied by increased resistance to other forms of denaturation, such as cleavage by proteases and chemical denaturation by guanidine hydrochloride or urea. A similar trend is seen in the evolved thermostable esterases. Thermostable 8G8 is more resistant than wild type to cleavage by trypsin and to denaturation by guanidine hydrochloride (Gershenson *et al.,* 2000).

In this directed evolution experiment, increased thermostability was accompanied by an increase in catalytic activity at all temperatures (Fig. 6), as measured on the substrate used during screening, *p*-nitrophenyl acetate. Variant 8G8 is 3.7 times more active than wild type at the wild-type T_{opt}; the improvement is 4.5-fold at 8G8's new T_{opt} of 60°C. Activity was allowed to decrease in the first generation (variant 1A5D1 was an intermediate product of a separate evolution experiment to increase *p*-nitrobenzyl esterase activity in aqueous organic solvent and differs from wild type at five positions), but was recovered in later generations.

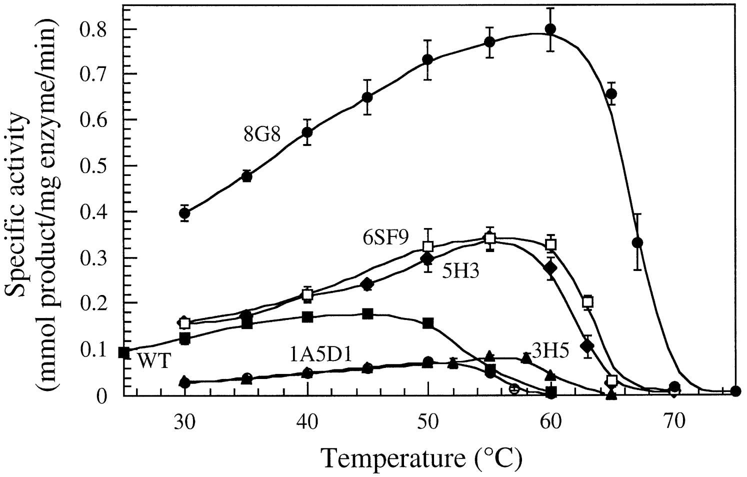


FIG. 6. Dependence of activity on temperature for wild-type *p*NB esterase and thermostabilized mutants from the first, third, fifth, sixth, and eighth generations.

Library data give useful information regarding the distribution of properties among the variants. Figure 7 shows a plot of room temperature activity versus stability for 1100 clones from the second generation *p*NB E library prepared by random mutagenesis. The ellipse represents values obtained for clones expressing the parent enzyme and show the inherent variability of the measurements, which is small compared to the distribution of values from the mutants. As expected, the majority of mutations are detrimental to both stability and activity. Furthermore, most stabilizing mutations come at the cost of catalytic activity (and *vice versa*). This does not mean, however, that catalytic activity and thermostability *always* come at the expense of the other, as we will discuss. It is to be expected that there are few single mutations that improve both traits simultaneously. Thus, the evolution of highly stable, highly active enzymes may need to take detours, or it may require the recombination of stabilizing and activating mutations.

A detailed analysis of the X-ray crystal structures of wild-type *p*NB E and thermostable 8G8 has been carried out (Spiller *et al.,* 1999). Detailed stabilization mechanisms for the various mutations are presented in the chapter by Orencia, Hanson, and Stevens in this volume; we will only

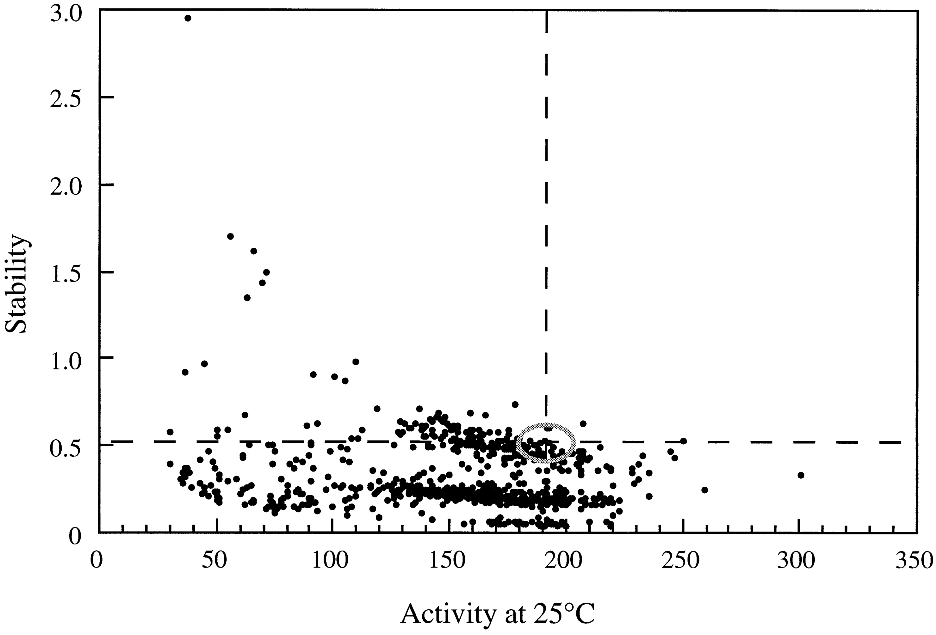


FIG. 7. Distribution of activity and stability in a library of random *p*NB esterase mutants.

briefly describe key results of that study. Stabilizing mutations are distributed throughout the structure (Fig. 8). Two surface loops composed of residues 66–74 and 414–420 do not appear in the wild-type crystal structure, but do appear in the structure of 8G8, indicating that stabilization was accompanied by reduced mobility in these regions. However, this reduction in mobility did not compromise catalytic efficiency, as the stable enzyme is also considerably more active. The connection between stability, mobility, and catalytic activity is thus highly complex. Finally, although two of the four mutations that increase catalytic activity (Leu313⇨Phe and Ala400⇨Thr) are located in the active site, the remaining two are far from it. The propagation of subtle effects over long distances is a common theme of molecular solutions obtained by directed evolution.

The dynamic properties of wild-type *p*NB E and several thermostable variants were investigated by monitoring the phosphorescence lifetimes of buried tryptophan residues at room temperature (Gershenson *et al.,*

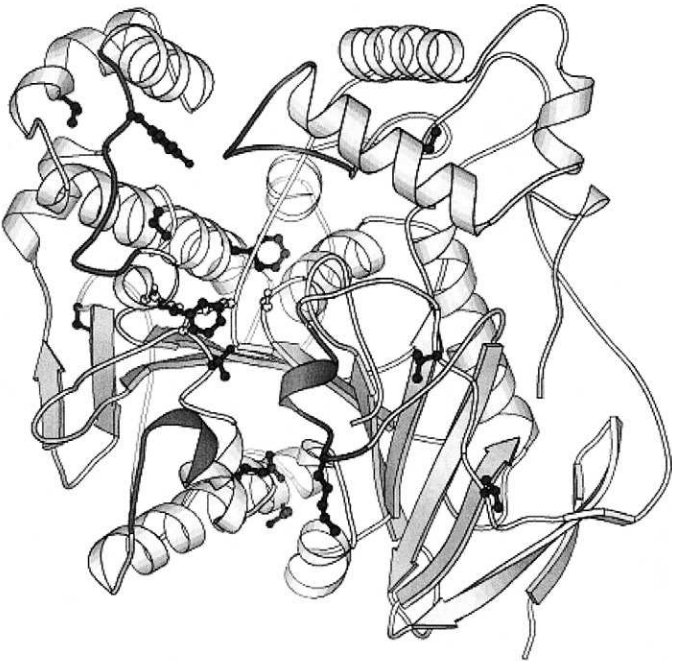


FIG. 8. MolScript™ (Kraulis, 1991) diagram of the three-dimensional structure of *p*NB esterase variant 8G8 (Spiller *et al.,* 1999). Mutated residues are shown in black ball-and-stick. Catalytic residues are shown in white ball-and stick. Black portions indicate stabilized loop regions.

2000). In the absence of oxygen (which effectively quenches Trp phosphorescence), the primary mechanism for phosphorescence decay is vibrational coupling between the triplet and ground states due to out-of-plane distortions of the aromatic ring. Phosphorescence lifetimes depend on the flexibility of the Trp side chain on the picosecond to nanosecond time scale, and are thus a function of the Trp residue's local environment. Trp residues located in highly ordered regions will undergo fewer out-of-plane fluctuations, and hence have longer phosphorescence lifetimes. Results show that wild-type *p*NB E has shorter lifetimes than any of the five variants studied, indicating reduced conformational motions in the local Trp environments in the stabilized variants. However, increases in melting temperature are not always accompanied by increases in phosphorescence lifetimes (Fig. 9). In two cases, significant increases in T_m were accompanied by decreased lifetimes (3H5 and 6SF9). Thus increases in thermal stability can be accompanied by *increased* flexibility, at least in the vicinity of the phosphorescing Trp residues. As mentioned, several loop regions that are poorly defined in the wild-type crystal structure are well defined in the variant 1A5D1 and all subsequent variants. Interestingly, the phosphorescing Trp residues, which are not located close to either the loops themselves or the stabilizing mutations, show increased lifetimes in the mutant 1A5D1 relative

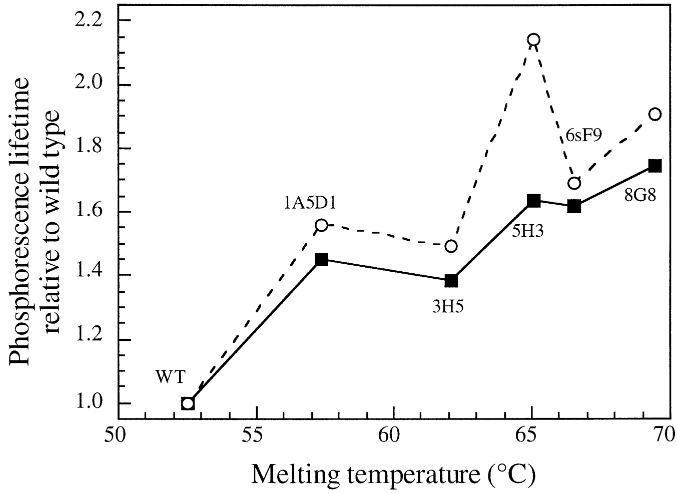


FIG. 9. Dependence of phosphorescence lifetimes on thermostability in the family of laboratory-evolved *p*NB esterases. Open circles indicate the long lifetime component of the phosphorescence signal, and filled squares indicate the short lifetime component (see Gershensen *et al.*, 2000).

to wild type. This suggests that stabilization involves long-range effects that propagate throughout the structure.

Also of interest is the relationship between phosphorescence lifetimes and catalytic activity. 1A5D1 and 3H5 are both less active toward pNPA than wild type, while 6sF9 and 8G8 are more active (Fig. 6). If flexibility near the phosphorescing Trp residues were monitoring motions critical to catalytic activity, then one would expect that 1A5D1 and 3H5 would have the longest lifetimes and 8G8 would have the shortest lifetimes, with wild type and 6sF9 somewhere in between. Instead, lifetimes do not appear to be correlated with activity. It is possible to reduce conformational motions in particular regions of the protein without interfering with motions required for efficient catalysis.

2. *Subtilisin E*

The subtilisins are a large family of serine proteases that have been extensively studied because of their importance in the detergent industry. The retain a common conserved fold and identical active site residues (Fig. 10). Two conserved calcium-binding sites are present in all subtilisins, and additional sites may be present in subtilisins from different organisms (Siezen and Leunissen, 1997).

Subtilisin E from *Bacillus subtilis* is a mesophilic enzyme whose mature sequence consists of 275 amino acids. Subtilisin E was evolved so that its thermostability and T_{opt} were identical to those of its thermophilic homolog thermitase (Zhao and Arnold, 1999). Because subtilisin denaturation is irreversible, stability was assayed by measuring residual activity after incubation at elevated temperatures. As with *p*NB E, thermostable mutants that retained the most catalytic activity at room temperature were selected. Five generations of random mutagenesis/recombination and screening generated subtilisin E 5-3H5, whose eight thermostabilizing amino-acid substitutions increased the enzyme's half-life at 65°C more than 200-fold (Table I) and raised its T_{opt} by 17°C (Fig. 11). The inactivation rate of the evolved enzyme at 83°C is indistinguishable from that of thermitase, while wild type loses 100% of its activity almost immediately at this temperature. As with *p*NB E, increased thermostability was accompanied by an increase in activity at all temperatures.

It was possible to identify each of the amino-acid substitutions as contributing either to activity, to stability, to both, or to neither. The mutations reveal that multiple small improvements and intramolecular interactions work together to increase stability through a variety of strategies. In this, the directed evolution results support the conclusions that have emerged from comparisons of related enzymes from mesophiles

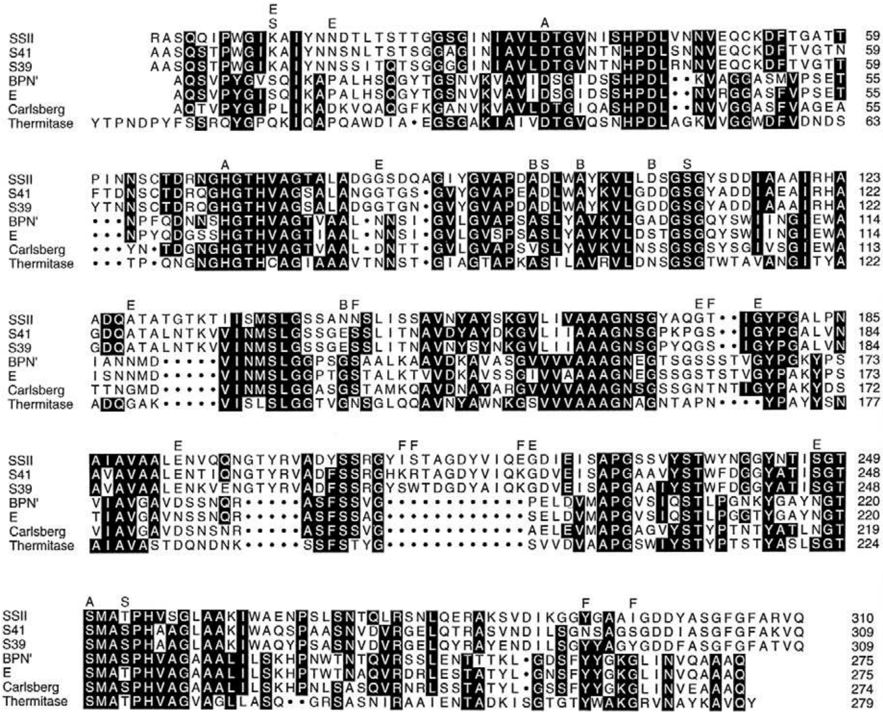

E
S E A
SSII RASQQIPWGIKAIYNNDTLTSTTGGSGINIAVLDTGVNISHPDLVNNVEQCKDFTGATT 59
S41 AASQSTPWGIKAIYNNSNLTSTSGGAGINIAVLDTGVNTNHPDLSNNVEQCKDFTVGTN 59
S39 AASQSTPWGIKAIYNNSSITQTSGGGGINIAVLDTGVNTNHPDLRNNVEQCKDFTVGTT 59
BPN' AQSVPYGVSQIKAPALHSQGYTGSNVKVAVIDSGIDSSHPDL••KVAGGASMVPSET 55
E AQSVPYGISQIKAPALHSQGYTGSNVKVAVIDSGIDSSHPDL••NVRGGASFVPSET 55
Carlsberg AQTVPYGIPLIKADKVQAQGFKGANVKVAVLDTGIQASHPDL••NVVGGASFVAGEA 55
Thermitase YTPNDPYFSSRQYGPQKIQAPQAWDIA•EGSGAKIAIVDTGVQSNHPDLAGKVVGGWDFVDNDS 63
A E BS B B S
SSII PINNSCTDRNGHGTHVAGTALADGGSDQAGIYGVAPDADLWAYKVLLDSGSGYSDDIAAAIRHA 123
S41 FTDNSCTDRQGHGTHVAGSALANGGTGS•GVYGVAPEADLWAYKVLGDDGSGYADDIAEAIRHA 122
S39 YTNNSCTDRQGHGTHVAGSALADGGTGN•GVYGVAPDADLWAYKVLGDDGSGYADDIAAAIRHA 122
BPN' •••NPFQDNNSHGTHVAGTVAAL•NNSI•GVLGVAPSASLYAVKVLGADGSGQYSWIINGIEWA 114
E •••NPYQDGSSHGTHVAGTIAAL•NNSI•GVLGVSPSASLYAVKVLDSTGSGQYSWIINGIEWA 114
Carlsberg •••YN•TDGNGHGTHVAGTVAAL•DNTT•GVLGVAPSVSLYAVKVLNSSGSGSYSGIVSGIEWA 113
Thermitase •••TP•QNGNGHGTHCAGIAAAVTNNST•GIAGTAPKASILAVRVLDNSGSGTWTAVANGITYA 122
E BF EF E
SSII ADQATATGTKTIISMSLGSSANNSLISSAVNYAYSKGVLIVAAAGNSGYAQGT••IGYPGALPN 185
S41 GDQATALNTKVVINMSLGSSGESSLITNAVDYAYDKGVLIIAAAGNSGPKPGS••IGYPGALVN 184
S39 GDQATALNTKVVINMSLGSSGESSLITNAVNYSYNKGVLIIAAAGNSGPYQGS••IGYPGALVN 184
BPN' IANNMD•••••VINMSLGGPSGSAALKAAVDKAVASGVVVVAAAGNEGTSGSSSTVGYPGKYPS 173
E ISNNMD•••••VINMSLGGPTGSTALKTVVDKAVSSGIVVAAAAGNEGSSGSTSTVGYPAKYPS 173
Carlsberg TTNGMD•••••VINMSLGGASGSTAMKQAVDNAYARGVVVVAAAGNSGSSGNTNTIGYPAKYDS 172
Thermitase ADQGAK•••••VISLSLGGTVGNSGLQQAVNYAWNKGSVVVAAAGNAGNTAPN••••YPAYYSN 177
E FF FE E
SSII AIAVAALENVQQNGTYRVADYSSRGYISTAGDYVIQEGDIEISAPGSSVYSTWYNGGYNTISGT 249
S41 AVAVAALENTIQNGTYRVADFSSRGHKRTAGDYVIQKGDVEISAPGAAVYSTWFDGGYATISGT 248
S39 AVAVAALENKVENGTYRVADFSSRGYSWTDGDYAIQKGDVEISAPGAAIYSTWFDGGYATISGT 248
BPN' VIAVGAVDSSNQR•••••ASFSSVG••••••••••••••PELDVMAPGVSIQSTLPGNKYGAYNGT 220
E TIAVGAVNSSNQR•••••ASFSSAG••••••••••••••SELDVMAPGVSIQSTLPGGTYGAYNGT 220
Carlsberg VIAVGAVDSNSNR•••••ASFSSVG••••••••••••••AELEVMAPGAGVYSTYPTNTYATLNGT 219
Thermitase AIAVASTDQNDNK•••••SSFSTYG••••••••••••••SVVDVAAPGSWIYSTYPTSTYASLSGT 224
A S F F
SSII SMATPHVSGLAAKIWAENPSLSNTQLRSNLQERAKSVDIKGGYGAAIGDDYASGFGFARVQ 310
S41 SMASPHAAGLAAKIWAQSPAASNVDVRGELQTRASVNDILSGNSAGSGDDIASGFGFAKVQ 309
S39 SMASPHAAGLAAKIWAQYPSASNVDVRGELQYRAYENDILSGYYAGYGDDFASGFGFATVQ 309
BPN' SMASPHVAGAAALILSKHPNWTNTQVRSSLENTTTKL•GDSFYYGKGLINVQAAAQ 275
E SMATPHVAGAAALILSKHPTWTNAQVRDRLESTATYL•GNSFYYGKGLINVQAAAQ 275
Carlsberg SMASPHVAGAAALILSKHPNLSASQVRNRLSSTATYL•GSSFYYGKGLINVEAAAQ 274
Thermitase SMATPHVAGVAGLLASQ••GRSASNIRAAIENTADKISGTGTYWAKGRVNAYKAVQY 279

FIG. 10. Sequence alignment of subtilisins S41, SSII, S39, BPN′, E, Carlsberg, and thermitase. Thermitase is an homologous subtilisin-like protease from the thermophilic bacterium *Thermoactinomyces vulgaris.* Residues conserved in four or more of the sequences are shaded. The positions of mutations discovered during the directed evolution of the various subtilisins are indicated above the alignment. E–subtilisin E, F–subtilisin S41, S–subtilisin SSII, B–subtilisin BPN′. Active site residues are indicated (A).

TABLE I

Stability and Activity Parameters for Wild-type Subtilisin E and 5-3H5[a]

	Half-life (65°C) (min)	k_{cat} (s^{-1})	K_M (mM)	k_{cat}/K_M (s^{-1} mM^{-1})
Wild type	4.9	25.4	0.385	66.0
5-3H5	1030	55.8	0.151	373.0

[a] Activity parameters were determined at 37°C, pH 8.0 against the substrate *N*-succinyl-Ala-Ala-Po-Phe-*p*-nitroanilide. Half-lives were measured at pH 8.0, 1 mM $CaCl_2$ (Zhao and Arnold, 1999). Errors are ±10%.

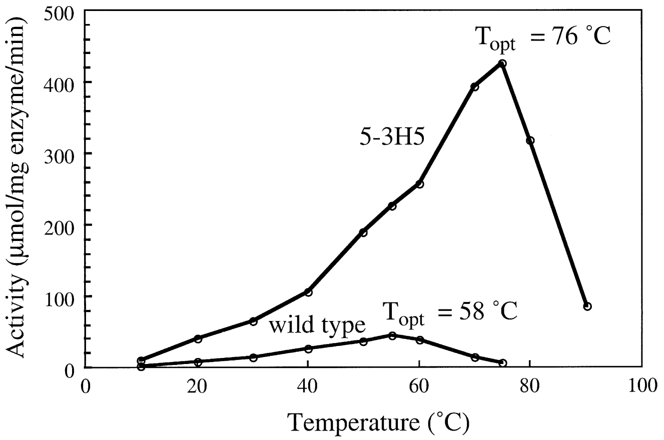


FIG. 11. Dependence of activity on temperature for wildtype subtilisin E and the thermostable mutant 5-3H5.

and thermophiles. Specific mutations and stabilization strategies are discussed in Section IV,D.

At least several hundred members of the subtilisin-like protease family have been identified, and sequence alignments indicate great sequence variability within the family (Siezen and Leunissen, 1997). All the mutations that stabilize subtilisin E are found in at least one known member of the subtilisin family (Fig. 10). Of the eight thermostabilizing mutations identified by Zhao and Arnold, five are found exclusively in mesophilic subtilisins, while the remaining three are found in both mesophilic and thermophilic subtilisins. Sequence comparisons of mesophilic and thermophilic subtilisins would thus not reveal the stabilizing potential of the mutations discovered by directed evolution. Subtilisin E and thermitase differ at 157 amino acid positions, yet only eight mutations were sufficient to convert subtilisin E into a functional equivalent of thermitase. Only two of these mutations are found in thermitase. Laboratory evolution took a quite different route to achieving high thermostability than nature has done with thermitase. This may reflect the unnaturally strong selective pressures that were placed on subtilisin E during the evolution experiment. It may also reflect different evolutionary trajectories for the two enzymes; thermitase, for example, may not have evolved from a mesophilic precursor. Even in the absence of these factors, however, we would expect that, given the vast number of possible sequences available to a protein the size of subtilisin E, multiple routes to high stability exist.

Molecular dynamics simulations of wild-type subtilisin E and the thermophile 5-3H5 at room temperature (300°K) revealed that the proteins have similar flexibilities (as estimated by the instantaneous RMS fluctuations about the average structure) (Colombo and Merz, 1999). The two enzymes also have comparable activities at room temperature. A more surprising result was found from simulations at 350°K. Thermostable 5-3H5 showed *greater* overall flexibility than wild type at this temperature, but at the same time retained more native-like structure. Wild type showed lower flexibility in terms of fluctuations about the average structure, but larger RMS deviations from the crystal structure (Figs. 12 and 13). In particular, two loop regions—residues 96-105 and 158-162—underwent motions that exposed side chains to solvent and allowed solvent penetration into the protein interior. Such events have been suggested as important to the onset of protein unfolding. These results point to the importance of specifying what is meant by ‘‘flexibility.’’ In terms of RMS fluctuations about the average structure, 5-3H5 is more flexible than wild type and, far from being destabilizing, this increased conformational entropy is probably responsible for stabilizing the native state. In terms of deviations from the crystal structure (which could also be considered a measure of flexibility), wild type is more ‘‘flexible’’ than 5-3H5. This increased flexibility, particularly in the two loop regions, may favor the initiation of unfolding.

In order to function, enzymes must maintain their active sites, particularly their catalytic residues, in the correct orientation for catalysis. For subtilisin E, the key catalytic residues are Ser221, His64, and Asp32. In both the low and high temperature simulations, 5-3H5 retained the active site geometry. In addition, at high temperature Ser221 showed smaller deviations from the crystal structure in 5-3H5 than in wild type, even though average flexibility was greater. Thus results for the active site residues are similar to the results for the protein overall.

3. Subtilisin S41

In another set of experiments, directed evolution was used to stabilize a psychrophilic subtilisin, S41, isolated from the Antarctic bacterium TA41 (Miyazaki *et al.,* 2000; Miyazaki and Arnold, 1999). S41 conserves the overall subtilisin fold and shares relatively high identity with other subtilisins (Fig. 10), but contains several features not found in the common mesophilic subtilisins, including a highly charged surface, several inserted surface loops, and decreased numbers of salt bridges and aromatic-aromatic interactions (Davail *et al.,* 1994). In common with other enzymes from psychrophilic sources, S41 is less stable at high

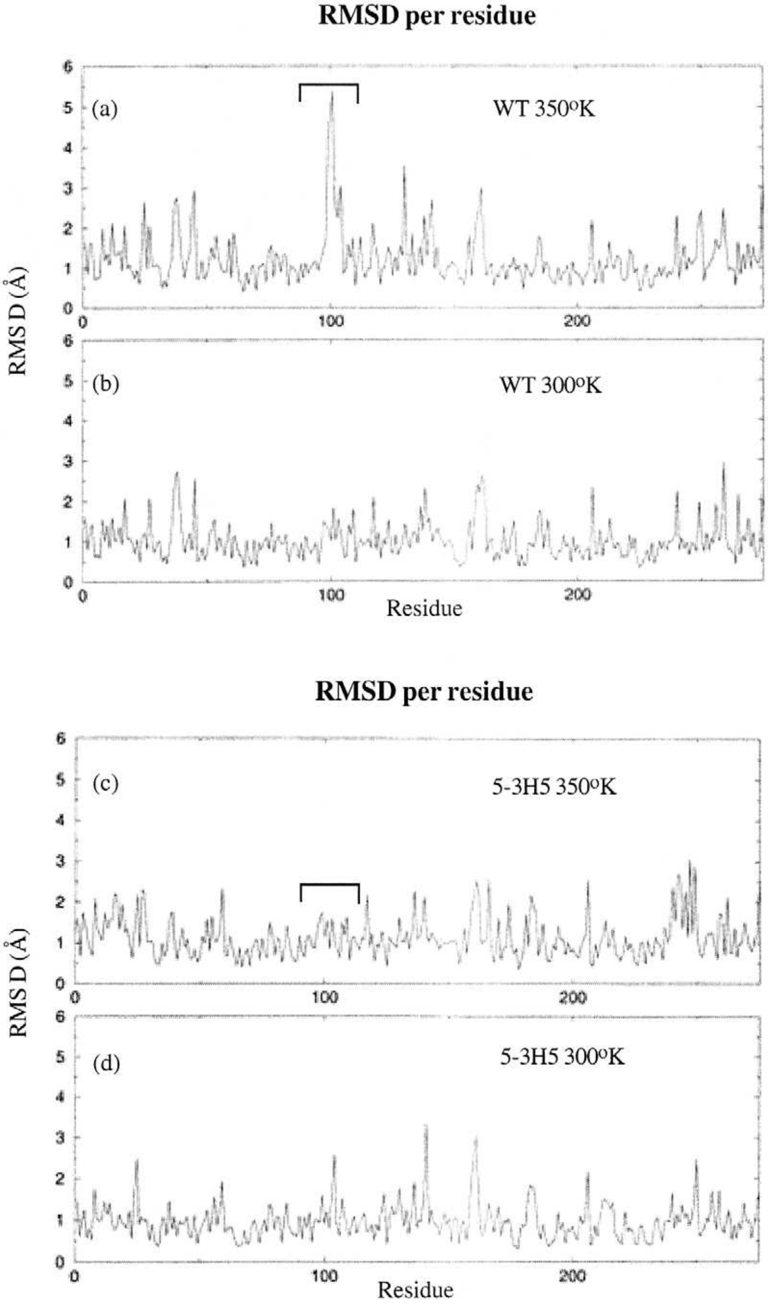


FIG. 12. RMS deviations per residue from the crystal structure during molecular dynamics simulations of (a) subtilisin E at 350°K, (b) subtilisin E at 300°K, (c) 5-3H5 at 350°K, (d) 5-3H5 at 300°K (Colombo and Merz, 1999). Bars indicate the loop region 96–105.

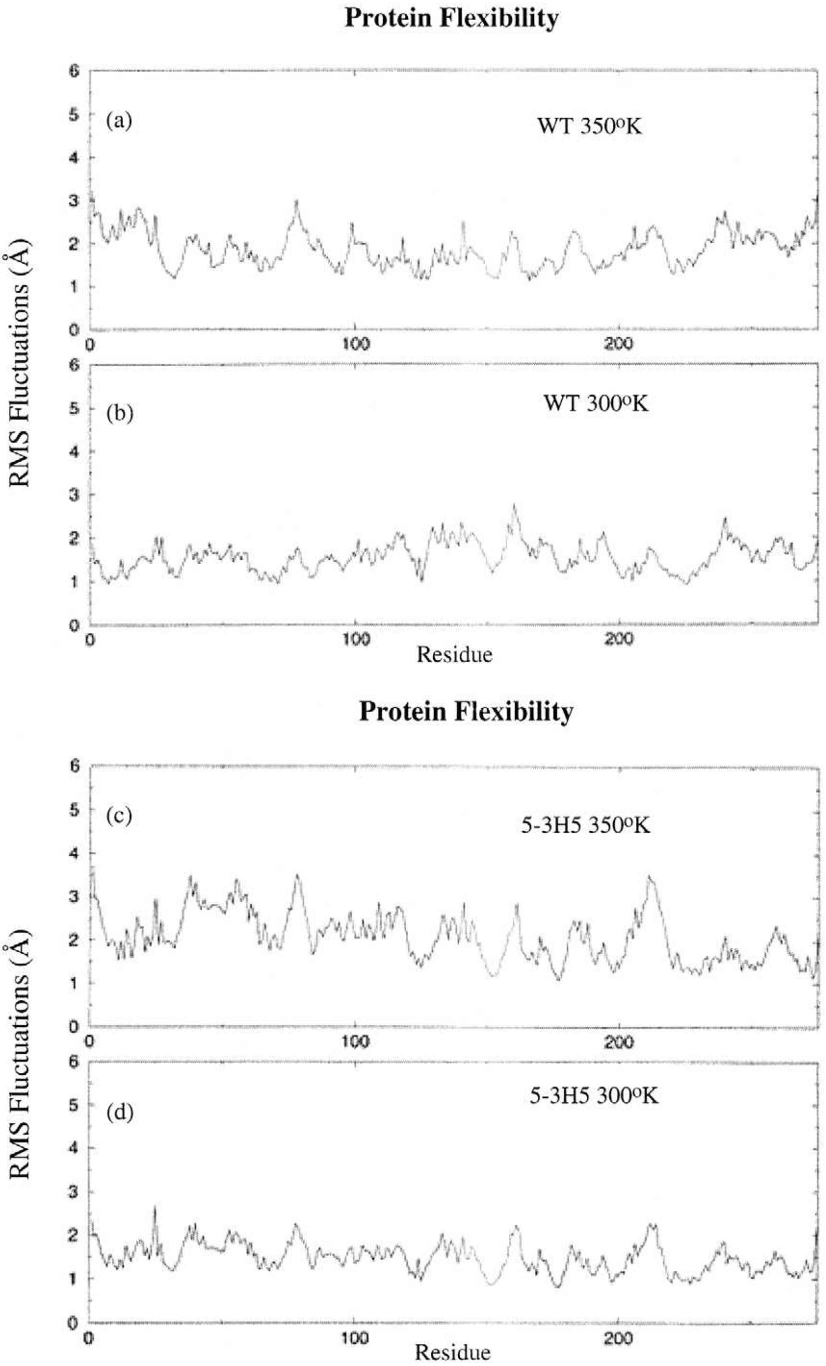


FIG. 13. RMS fluctuations per residue about the time-averaged structure during molecular dynamics simulations of (a) subtilisin E at 350°K, (b) subtilisin E at 300°K, (c) 5-3H5 at 350°K, (d) 5-3H5 at 300°K (Colombo and Merz, 1999).

temperatures and more active at low temperatures (<15°C) than its mesophilic counterparts.

The plot of the stabilities and activities of clones from the first generation S41 random mutant library shows once again that most mutations are detrimental to stability and activity (Fig. 14). However, compared to the esterase library (Fig. 7), there are more mutants with improvements in both properties, suggesting that the two enzymes have different adaptive potentials. This may be due to the relatively poor stability of S41, or it may reflect constraints intrinsic to the three-dimensional structures of the two proteins. Evidence for the former can be found by comparing the results for the first generations of the psychrophilic subtilisin S41 and the mesophilic subtilisin E. Screening 864 mutants of S41 yielded nine thermostabilized variants (a "hit rate" of approximately 1%) (Miyazaki and Arnold, 1999); in contrast, screening 5000 subtilisin E mutants identified five thermostable variants (a "hit rate" of only 0.1%) (Zhao and Arnold, 1999).

Of the nine thermostabilized S41 variants identified in the first generation, three had substitutions at residue 211, and one contained a substitution at 212. Saturation mutagenesis was performed at both sites in order to explore a greater sequence diversity than that accessible through error-prone PCR (Miyazaki and Arnold, 1999). One hundred and five of the clones from the library that showed reasonable activity were assayed

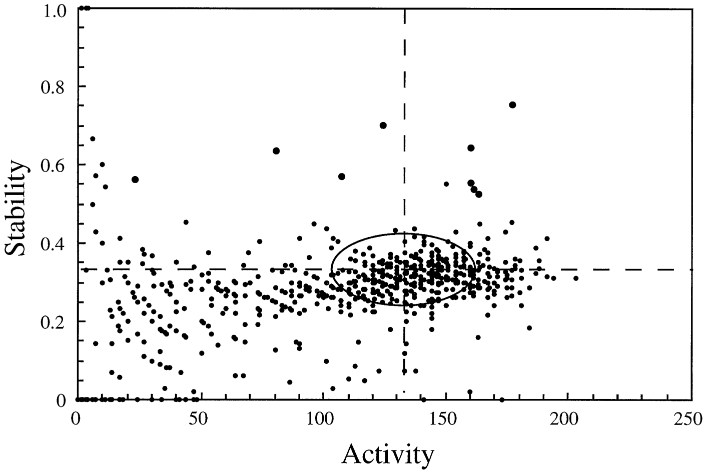


FIG. 14. Distribution of activity and stability in a first generation random mutant library of S41. Ellipse shows one standard deviation of the values obtained for wild-type clones in this assay.

for thermostability. The stabilities were distributed over a wide range, with a large fraction (~40%) that were more stable than wild-type S41. The most stable variant found in the saturation mutagenesis library was considerably more stable (half-life at 60°C, 84 minutes) than the best variant found in the error-prone PCR library (half-life at 60°C, 27 minutes).

Sequencing of the four most stable variants from the saturation mutagenesis library revealed the following residues at 211/212: Trp/Ser, Pro/Val, Leu/Val, and Pro/Ala. In all cases, the sites were occupied by hydrophobic residues quite different from the wild-type sequence Lys/Arg. All these substitutions require multiple base changes, and are therefore not accessible by error-prone PCR. Additionally, all were nonconservative according to the Dayhoff 250PAM matrix (Dayhoff *et al.,* 1978). The evolution of thermostability was thus accelerated by access to nonconservative mutations (Miyazaki and Arnold, 1999).

Recombination of the improved first generation mutants followed by another round of random mutagenesis and screening yielded S41 3-2G7, with a half-life at 60°C that is ~500 times that of wild type (Miyazaki *et al.,* 2000) (Fig. 15). 3-2G7 contains seven amino acid substitutions, all of which are thermostabilizing (listed and described in Section IV,D). Activity and stability parameters for wild-type S41 and 3-2G7 are given in Table II. As was found for subtilisin E and *p*NB E, the stabilized S41 is more active than wild type over the entire temperature range, as measured on the substrate used during screening, s-AAPF-*p*NA (Fig. 15). Although 3-2G7 retains the efficient low temperature catalysis of wild type, it is even more stable than the mesophilic subtilisin BPN′. Further rounds of evolution have generated even more stable variants, so that psychrophilic S41 has been converted into a truly thermophilic enzyme (Miyazaki, K., unpublished results).

Similar to what was found for subtilisin E, all the amino-acid substitutions in the thermostable S41 variant 3-2G7 are found in other subtilisn sequences. Looking at the sequences of selected, closely related homologs (Fig. 10), we see that two of the seven stabilizing mutations (Ser175⇨Thr and Lys221⇨Glu) are present in the mesophilic SSII, and one (Ser175⇨Thr) is also present in subtilisins E and Carlsberg. Further, comparing only S41, S39, and SSII, we see that all substitutions occurred at positions that are variable among the three proteins. This demonstrates once again the difficulties of deducing stabilization strategies from sequence comparisons. As was seen in the case of subtilisin E, the stabilizing potential of individual amino acid substitutions depends on the background sequence into which they are introduced. A substitu-

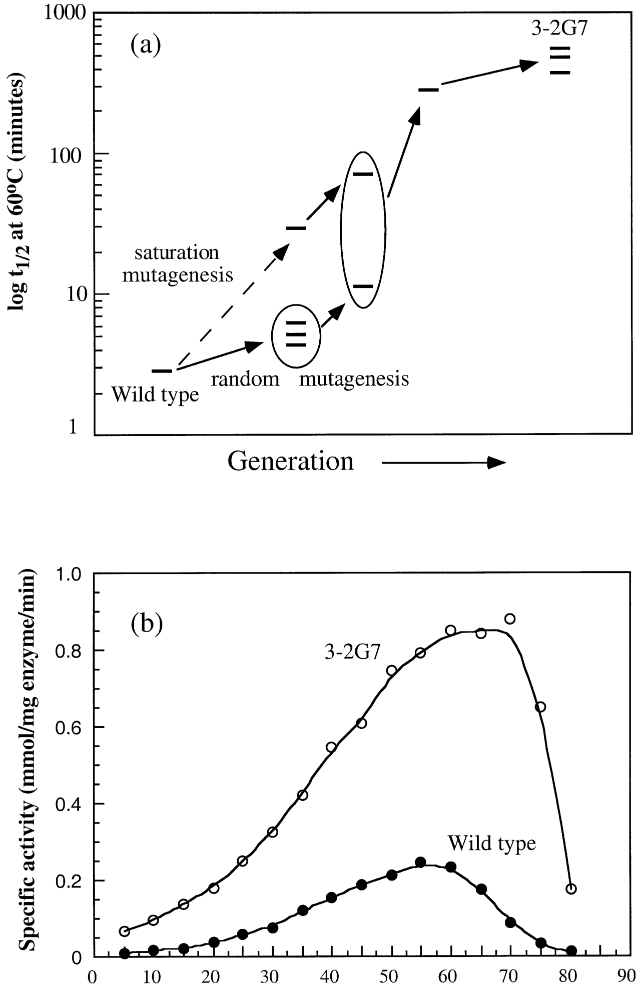


FIG. 15. Evolution of (a) stability (half-life at 60°C) and (b) activity for subtilisin S41.

tion that is highly stabilizing in the context of one sequence may be neutral or even deleterious in the context of another.

An important mechanism for stabilization in 3-2G7 is increased affinity for calcium ions. It has been noted (Feller and Gerday, 1997) that psychrophilic enzymes often exhibit lower affinity for metal ions than their mesophilic counterparts. The stabilization of proteins by calcium

Table II

Stability and Activity Parameters for Wild-type S41 and SSII and Evolved Mutants

	Half-life[a] (minutes)	k_{cat} (10°C) (s^{-1})	K_M (10°C) (mM)	k_{cat} (30°C) (s^{-1})	K_M (30°C) (mM)	k_{cat} (60°C) (s^{-1})	K_M (60°C) (mM)
S41	9.3	23.3	0.28	63.8	0.27	264	0.91
3-2G7	566	48.2	0.14	179	0.21	553	0.50
SSII	13.5	15.7	0.28	90	0.34	258	0.59
P3C9	4.1	104	0.19	258	0.25	579	0.53

Activity parameters were determined at pH 8.5 against the substrate *N*-succinyl-Ala-Ala-Po-Phe-*p*-nitroanilide. [a] Half-lives for S41 and 3-2G7 were determined at pH 8.5, 60°C; those for SSII and P3C9 were determined at pH 8.5, 70°C. Errors are ±10%.

binding, either through increasing the affinity of existing sites or creating new sites, is well documented (Voordouw *et al.,* 1976; Teplyakov *et al.,* 1990; Braxton and Wells, 1992). This strategy clearly has high potential for stabilization and was readily discovered by the evolution experiment. Increased affinity for calcium was also implicated in the stabilization of subtilisin E (Colombo and Merz, 1999), although direct experimental evidence is lacking. Interestingly, Narinx *et al.* (1998) discovered a very different strategy for stabilization by Ca^{2+} binding in subtilisin S39, which shares very high (88.3%) sequence identity with S41. They introduced a mutation, Thr 85⇨Asp, that dramatically increases the apparent K_a for Ca^{2+} and increases the enzyme half-life at 50°C by a factor of ten. When this same mutation is introduced into S41, however, it decreases rather than increases stability (Miyazaki, K., unpublished results). Thus, it is difficult to extrapolate results from one protein to another, even when they share high sequence identity.

Although a model structure of subtilisin S41 was constructed based on its homology to subtilisins of known structure (Miyazaki *et al.,* 1999), little detailed information on stabilization mechanisms could be gained from examining the positions of thermostabilizing mutations. Increased affinity for Ca^{2+} apparently plays a key role in stabilization, but it is not clear how this increased affinity was achieved. Based on sequence homology, S41 probably contains at least two Ca^{2+} sites that are found in other subtilisins: a high-affinity site that is essential for activity, and a lower-affinity site whose effect on stability depends strongly on the extent of Ca^{2+} binding. None of the mutations, however, are close to the predicted locations of these sites. More detailed experimental structural data will be required to explain the increased affinity.

It was suggested (Davail *et al.,* 1994) that the extended surface loops present in S41 but not in most of its mesophilic homologs are partly responsible for S41's low thermostability. The fact that very similar loops are also present in the closely similar mesophilic subtilisin SSII (see below), however, makes it clear that loops *per se* cannot be responsible. Loops can be weak points, possibly serving as initiation sites for unfolding. Positions 211 and 212, near the N-terminus of S41's largest loop (extending from position 210 to 221) were found to be particularly important for the thermostabilization of S41. Without structural data it is not possible to know the exact mechanism by which the most stabilizing substitutions at these positions, Lys211⇨Pro and Arg212⇨A1a, exert their stabilizing effects. However, the introduction of a Pro residue can rigidify the loop by restricting the number of available main chain conformations. Stabilization by the introduction of prolines into loop regions is well documented (Matthews, 1993; Watanabe *et al.,* 1994).

4. 3-Isopropylmalate Dehydrogenase

A successful example of the evolution of thermal stability by selection in a thermophilic host is provided by recent work on 3-isopropylmalate dehydrogenase (IPMDH) from *Bacillus subtilis* (Akanuma *et al.,* 1998). An additional example, not reviewed here, is kanamycin nucleotidyltransferase (Matsumura and Aiba, 1985; Liao *et al.,* 1986). IPMDH catalyzes the oxidative decarboxylation of (2R,3S)-3-isopropylmalate to 2-oxoisocaproate, an essential step in the leucine biosynthesis pathway. The gene encoding the mesophilic IPMDH from *Bacillus subtilis* was inserted into the chromosome of an IPMDH-deficient strain of the extreme thermophile *Thermus thermophilus.* Initially, the investigators relied on the appearance of spontaneous mutations to create the molecular diversity. Thermostabilized IPMDH variants were identified by growing the *T. thermophilus* host in the absence of leucine at 61°C, a temperature at which the wild-type IPMDH is no longer functional. Positive clones were subjected to a second selection at 66°C, and positives from this round were grown at 70°C.

Stabilizing mutation Thr308⇨Ile was discovered in the first selection, followed by Ile95⇨Leu in the second and Met292⇨Ile in the third generation. The final triple mutant retains 100% of its activity after 10 minutes at 50°C, and 50% of its activity after 10 minutes at 60°C. In contrast, wild type loses 50% of its activity after incubation at 50°C for 10 minutes. Although the mutant's activity at high temperatures (i.e., above the selection temperature) was substantially increased, its activity at 40°C remained essentially that of wild type. PCR-based random mutagenesis and selection was used to further stabilize the mutant IPMDH (Akanuma *et al.,* 1999). This produced two additional mutations (Thr22⇨Lys and Met256⇨Val) in a quintuple mutant that retains 80% of its activity after 10 minutes at 60°C and 50% after 10 minutes at 62°C. The k_{cat} of the quintuple mutant decreased ~2 fold at 40°C. This loss is entirely due to the Met256⇨Val substitution.

The three-dimensional structure of IPMDH from *B. subtilis* is not known, but the enzyme is highly similar to the IPMDH from *T. thermophilis,* whose structure is known. The protein exists as a dimer in which each subunit consists of two domains: one forms part of the dimer interface, and the other contains the N- and C-termini. The thermostabilizing mutations Thr308⇨Ile, Met292⇨Ile, Ile95⇨Leu are located in the region formed by the N- and C-termini, and not in the dimer interface. A study of the thermal unfolding of IPMDH (Iwasaki *et al.,* 1996) has shown that the domain formed by the termini unfolds before the domain containing the dimer interface. The mutations thus seem to have stabilized the enzyme by strengthening its weakest point. The Met256⇨Val substitution was in the interface and presumably increases stability by improving intersubunit interactions.

Aligned sequences of IPMDHs from *B. subtilis* and *T. thermophilis* are shown in Figure 16. Three of the five thermostabilizing mutations discovered by directed evolution are present in the natural thermophilic enzyme. The approach used to stabilize IPMDH differed in several ways from that used to evolve *p*NB E and subtilisins E and S41, in particular in that the selection explicitly required *activity* at high temperatures in addition to high thermostability. Unlike *p*NB esterase and the subtilisins, IPMDH was required not only to be active, but to be active within the context of the metabolic pathway of the host organism. Thus, even with the constraints imposed by biological function, mutations that increase thermal stability without sacrificing activity at lower temperatures can be discovered. It is important to note, however, that because IPMDH was *not* screened for activity at moderate temperature, a mutation detrimental to activity at lower temperatures (Met256⇨Val) was acquired.

D. *Directed Evolution of Low-Temperature Activity*

Psychrophilic enzymes have attracted significantly less attention than thermophilic ones, and there are correspondingly fewer examples of the directed evolution of psychrophilic properties. The studies carried out so far have focused on subtilisin proteases to generate higher activity at low temperature (Taguchi *et al.,* 1998; Wintrode *et al.,* 2000). These two studies present an instructive contrast: the target enzymes are very similar, but the mutagenesis and screening methods used and the results obtained differ substantially.

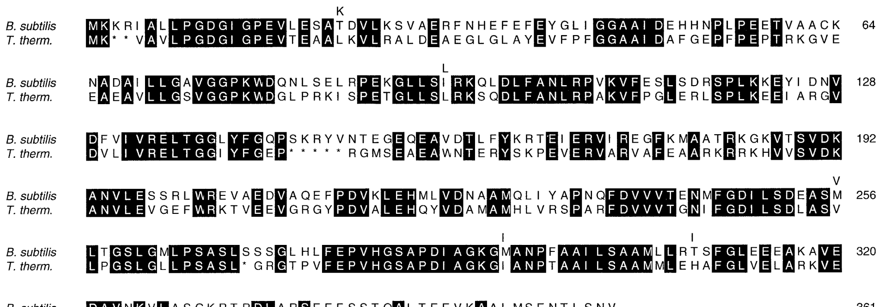

K
B. subtilis MKKRIALLPGDGIGPEVLESATDVLKSVAERFNHEFEFEYGLIGGAAIDEHHNPLPEETVAACK 64
T. therm. MK**VAVLPGDGIGPEVTEAALKVLRALDEAEGLGLAYEVFPFGGAAIDAFGEPFPEPTRKGVE
L
B. subtilis NADAILLGAVGGPKWDQNLSELRPEKGLLSIRKQLDLFANLRPVKVFESLSDRSPLKKEYIDNV 128
T. therm. EAEAVLLGSVGGPKWDGLPRKISPETGLLSLRKSQDLFANLRPAKVFPGLERLSPLKEEIARGV
B. subtilis DFVIVRELTGGLYFGQPSKRYVNTEGEQEAVDTLFYKRTEIERVIREGFKMAATRKGKVTSVDK 192
T. therm. DVLIVRELTGGIYFGEP*****RGMSEAEAWNTERYSKPEVERVARVAFEAARKRRKHVVSVDK
V
B. subtilis ANVLESSRLWREVAEDVAQEFPDVKLEHMLVDNAAMQLIYAPNQFDVVVTENMFGDILSDEASM 256
T. therm. ANVLEVGEFWRKTVEEVGRGYPDVALEHQYVDAMAMHLVRSPARFDVVVTGNIFGDILSDLASV
I
I
B. subtilis LTGSLGMLPSASLSSSGLHLFEPVHGSAPDIAGKGMANPFAAILSAAMLLRTSFGLEEEAKAVE 320
T. therm. LPGSLGLLPSASL*GRGTPVFEPVHGSAPDIAGKGIANPTAAILSAAMMLEHAFGLVELARKVE
B. subtilis DAVNKVLASGKRTRDLARSEEFSSTQAITEEVKAAIMSENTISNV 361
T. therm. DAVAKAL*LETPPPDLGGSAGTEAFTATVLRHLA

FIG. 16. Sequence alignment for IPMDH from *Bacillus subtilis* and *Thermus thermophilis.* Amino-acid substitutions discovered by directed evolution are shown above the alignment.

1. *Subtilisin SSII*

The mesophilic subtilisin-like protease SSII, isolated from the tropical bacterium *Bacillus sphaericus* (Wati *et al.,* 1997), shares high sequence identity with S41 (77.4%). SSII contains the same inserted loop regions found in S41 (see Fig. 10). Furthermore, SSII contains fourteen of the twenty-one Asp residues present in S41, in addition to five Asp residues not present in S41. Despite these similarities, however, SSII displays no particular psychrophilic characteristics. Its activity at low temperatures and its stability at high temperatures are comparable to those of mesophilic subtilisins such as BPN′ (Wintrode *et al.,* 2000). The high degree of sequence similarity, combined with their different functional properties, made S41 and SSII attractive for the study of cold adaptation. Directed evolution successfully increased the thermostability of subtilisin S41 while retaining its high catalytic activity at low temperatures (see Section IV,C,3). In parallel with this effort, SSII was evolved into a psychrophilic enzyme. This dual approach allowed exploration of the relationship between stability, low temperature activity, and selective pressure in two proteins that share highly similar sequences and, presumably, overall fold.

During the evolution of S41, selective pressure was applied to both stability and activity. Stabilized mutants were accepted only if they showed no or little decrease in activity. Wintrode *et al.* (2000) adopted a different strategy for the evolution of SSII's activity at low temperature. Random mutant libraries of SSII were prepared either by random mutagenesis or *in vitro* recombination and screened for catalytic activity at 10°C. Libraries were *not* screened for thermostability, and mutants that showed improved activity at 10°C were selected regardless of changes, positive or negative, in stability. This experiment addressed the question of what happens to thermostability when no selective pressure is applied.

Three mutants with improved activity at 10°C were identified in the first generation library. Recombination of these resulted in variant (P3C9) whose k_{cat} at 10°C is 6.6 times that of wild-type SSII and 4.5 times that of the naturally cold-adapted S41 (Fig. 17). The mutant SSII k_{cat}/K_M is 9.6 times that of wild type and 6.6 times S41's (Table II). The activities of naturally psychrophilic enzymes are generally less temperature dependent than those of their mesophilic counterparts. The temperature dependence of enzyme activity is often expressed in terms of the increase in k_{cat} brought about by raising the reaction temperature by 10°C ("Q_{10}"). The Q_{10} of P3C9 is ~1.2, more than two times lower than wild-type SSII and comparable to the value of ~1.3 determined for S41.

Comparing the effect of evolution on the temperature dependence of k_{cat} for SSII and S41 reveals how strongly enzyme properties depend

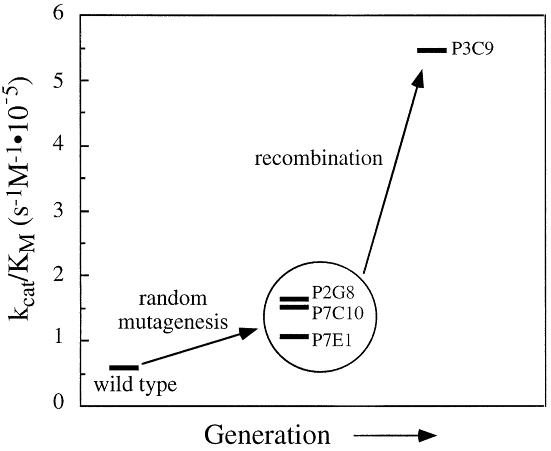


FIG. 17. Evolution of low temperature activity in subtilisin SSII.

on the temperature at which selective pressure is applied. SSII mutants were selected for increased activity at 10°C. Figure 18 shows that the relative superiority of P3C9 over wild-type is temperature dependent, with the increase greater at 10°C. The activities of S41 mutants were assayed at 25° to 30°C, and the greatest different in k_{cat} between wild type S41 and 3-2G7 was found at 30°C (Fig. 18). Interestingly, the relative superiority of the mutant S41 over the wild-type enzyme is much less than it is for SSII.

So what did happen to the thermostability of SSII during adaptation for activity in the cold? P3C9 is less thermostable than its mesophilic counterpart SSII, with a half-life of inactivation at 70°C that decreased approximately threefold. However, there is no strict inverse correlation between stability and low temperature activity during the evolution. While low temperature activity was improved, thermal stability both increased and decreased, although the overall trend was toward decrease (Fig. 19). This suggests that, rather than being inversely coupled, thermal stability is essentially decoupled from low-temperature activity. And, since most mutations are deleterious, stability decreases when not subject to any selective pressure.

Decreased affinity for stabilizing ligands such as Ca^{2+} has been identified as a factor that contributes to the poor thermostability of natural psychrophilic enzymes. This does not appear to be the case for the laboratory-evolved psychropile P3C9, which retained wild-type affinity for calcium (Wintrode *et al.,* 2000). The cold-activating mutations are distributed throughout the structure, as predicted using a homology-based model. There are substitutions at both surface and buried posi-

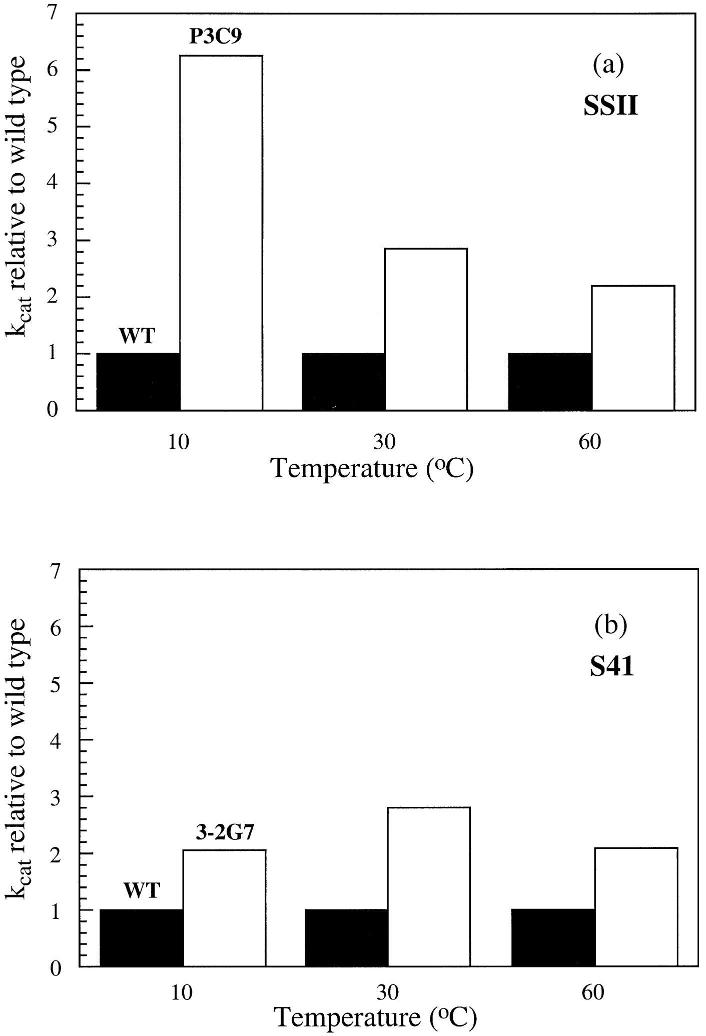


FIG. 18. (a) Rate constant k_{cat} of P3C9 relative to wild-type SSII at 10, 30, and 60°C. (b) k_{cat} of 3-2G7 relative to wild-type S41 at 10°, 30°, and 60°C.

tions, and both close to and far from the active site. None of the mutations is in or near the putative weak Ca^{2+} binding site, a finding that is consistent with the fact that P3C9 did not lose affinity for calcium.

2. *Subtilisin BPN′*

Taguchi *et al.* (1998) used directed evolution to increase the low-temperature activity of the mesophilic subtilisin BPN′. Random muta-

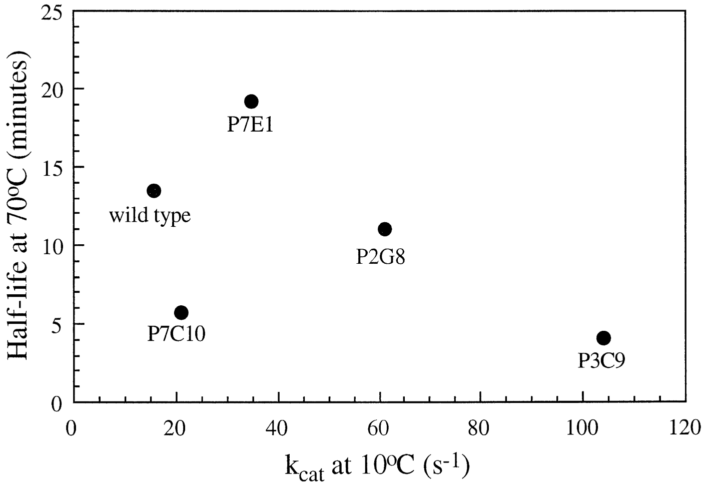


FIG. 19. Activity at 10°C versus stability at 70°C for wild-type SSII and evolved cold-active mutants.

tions were introduced into the gene coding for BPN′ by treatment with hydroxylamine. Mutant libraries were screened by growing transformed bacilli on petri plates containing 2% skim milk. The secreted proteases hydrolyze skim milk proteins, giving rise to a clear zone surrounding colonies secreting active protease. Transformed *bacteria* were grown on skim milk plates overnight at 37°C and then incubated at 10°C for two days. Improved mutants were identified by clear zone formation at 10°C.

The evolution was carried out in two stages. First, colonies were screened for activity-suppressed mutants (those with no detectable clear zone). These were then taken on to the next round of mutagenesis, and the resulting mutant library was screened for mutants with restored activity. This approach reduced the background of variants with wild-type levels of activity, allowing positive mutations to be identified more easily.

Screening yielded a single mutant enzyme, labeled m-63, that was twice as active against the substrate s-AAPF-*p*NA at 10°C. Subtilisin m-63 was only 50% more active than wild type at 37°C, and 40% more active at 50°C, indicating that activity was enhanced preferentially at low temperatures. The observed increase in specific activity was due entirely to a decrease in K_M, which, at 10°C, is 69 μM for m-63 compared to 135 μM for wild-type. This decrease in K_M may seem surprising, given the nature of the screen. Enzyme is secreted into a suspension containing 2% skim milk, which corresponds to a large excess of substrate. In the presence of excess substrate, enzymes should not, in general, be able to improve their specific activity through improvements in K_M. However, in this screen, proteins encounter substrate in the colloidal environment of the petri dish. Diffusion of both enzyme and substrate in this environment is substantially reduced compared to that in solution, and thus the effective concentration of substrate is smaller than the actual concentration. Under such conditions, the enzyme can improve its specific activity by decreasing K_M. This example demonstrates how the outcome of a directed evolution experiment depends on the details of the screen. The decrease in K_M is temperature dependent, with m-63's K_M preferentially lower than that of wild type at 10°C. A subsequent study on BPN′ using identical screening and mutagenesis methods generated mutant m-51, which was 70% more active than wild type at 10°C (Taguchi *et al.*, 1999). Unlike m-63, the increase in low temperature activity in m-51 was primarily due to an increase in k_{cat}.

Mutant m-63 contains three amino acid substitutions: Val 72⇨Ile, Ala 92⇨Thr, and Gly 131⇨Asp. Site-directed mutagenesis was used to evaluate the effect of each mutation. Val 72⇨Ile was identified as the initial activity-suppressing mutation: it caused a small decrease in k_{cat} and a substantial increase in K_M. Additionally, it resulted in a very large decrease in thermal stability as measured by the half-life at 60°C, reducing it from 205 minutes for wild type to less than 10 minutes. Individually, both Ala 92⇨Thr and Gly 131⇨Asp increased stability and activity, although the double mutant Ala 92⇨Thr, Gly 131⇨Asp was not as active as the triple mutant. The mechanism for the observed decrease in K_M is not known, although positions 92 and 131 are both close to the substrate binding region (Siezen and Leunissen, 1997).

Mutant m-51 also contained three substitutions: Ala88⇨Val, Ala98⇨Thr in the mature enzyme and Ala31⇨Thr in the pro-region. The suppressor mutation was found to be Ala31⇨Thr. None of the cold active variants in this study showed any loss of stability relative to wild type. This work demonstrates that activity can be increased at low temperatures, and that it can be done at no cost to thermostability.

E. Characteristics of Adaptive Mutations Discovered by Directed Evolution

The number of thermally adaptive mutations resulting from directed evolution studies is too small at present to support a detailed statistical analysis. Here we summarize some properties of the mutations discovered in the studies reviewed above, and compare them to the amino-acid differences seen among naturally occurring enzymes that have adapted to different temperatures. Lists of the amino-acid substitutions discovered during the work discussed above are given in Tables III and IV. The majority of adaptive mutations are located on the protein surface and are in loop or turn regions rather than in elements of regular secondary structure such as helices or sheets (Table V). In this, laboratory evolution results in trends similar to those seen in natural enzymes (Querol *et al.,* 1996). It is likely that this preference for surface sites reflects the differences in plasticity of the surface versus the core. A large body of work points to the importance of efficient interior packing in proteins to structural stability (Makhatadze and Privalov, 1995; Shakhnovich and Finkelstein, 1989; Dahiyat and Mayo, 1997). Single mutations in the interior are likely to disrupt this packing and lead to the loss of stability or folded structure. It is therefore expected that protein surfaces will be more tolerant to single mutations than proteins interiors. Similar considerations hold for the fact that mutations are most often found outside of helices or sheets. These regular secondary structures require specific geometries that are presumably more easily disrupted than loops. Few will argue that the surface is not tolerant to substitutions—the interesting result is that it offers such opportunity for adaptation. Different evolutionary experiments are likely to find different solutions to the same functional challenge. Higher-frequency random mutagenesis targeted to specific residues in the core, for example, may identify compensating mutation sets that provide superior thermostabilization.

Most of the mutations (64%–67%) are conservative in terms of hydrophobicity/hydrophilicity and charge. This is probably due to the combined effects of adaptation and the structure of the genetic code. Substi-

TABLE III

Amino-acid Substitutions Discovered During the Directed Evolution of Thermostability

*p*NB esterase	Subtilisin E	Subtilisin S41	IPMDH
Ala56Val	Pro14Leu	Ser145Ile	Thr22Lys
Ile60Val	Asn76Asp	Ser175Thr	Ile95Leu
Thr73Lys	Asn118Ser	Lys211Pro	Met256Val
Leu144Met	Ser161Cys	Arg212Ala	Met292Ile
Leu313Phe	Gly166Arg	Lys221Glu	Thr308Ile
His322Tyr	Asn181Asp	Asn291Ile	
Ala343Val	Ser194Pro	Ser295T	
Met358Val	Asn218Ser		
Tyr370Phe			
Ala400Thr			
Gly412Glu			
Ile437Thr			
Thr459Ser			

TABLE IV
Amino-acid Substitutions Discovered During the Directed Evolution of Low Temperature Activity

Subtilisin SSII	Subtilisin BPN′
Lys11Arg	Ala31Thr (pro region)
Asp98Asn	Val72Ile
Ser110Phe	Ala88Val
Thr253Ala	Ala92Thr
	Ala98Thr
	Gly131Asp

tutions that change polarity or charge may be more likely than conservative substitutions to disrupt protein structure and stability (particularly if they are located in the core), and are thus less likely to be found in functional mutants selected from random libraries. Additionally, and probably most important, nonconservative mutations are much less frequent in the library because of the distribution of codons in the genetic code. The genetic code is structured in such a way that single base substitutions, when they are not synonymous, are more likely to result in amino-acid substitutions that conserve important properties such as hydrophobicity and charge. The bias of PCR mutagenesis for transition

TABLE V
Characteristics of Thermostabilizing Mutations in p*NB Esterase, Subtilisin E, and Subtilisin S41*

Total number mutations	28
Conservative mutations (hydrophobic : hydrophilic)	18
Nonconservative mutations (hydrophobic : hydrophilic)	10
Conservative mutations (uncharged: charged)	19
Nonconservative mutations (uncharged : charged)	9
Conservative mutations (according to the Dayhoff matrix)	14
Nonconservative mutations (according to the Dayhoff matrix)	14
Surface	21
Buried	7
Located in loops, turns, or coils	22
Located in helices or sheets	6

mutations over transversions also limits access to nonconservative amino acid substitutions.

It is not yet clear whether any rules for thermal adaptation are to be found in the qualitative features of the mutations discovered during directed evolution. The trends seen in Table V (preference for surface substitutions over buried substitutions, conservative over nonconservative, etc.) would most likely be seen in any pool of functional mutants selected from random mutant libraries.

In considering the relationship of temperature-adaptive mutations to the sequences of naturally related homologous proteins, we will focus on the subtilisin proteases since they comprise a well characterized family of which a large number of sequences are known, including sequences from organisms living at a variety of temperatures. Here, we can begin to distinguish separate trends for thermostabilizing mutations and cold-activating mutations. Comparing results from the laboratory evolution of subtilisins E and S41 with those for BPN′ and SSII, we see that thermostabilizing mutations tend to occur at variable sites, while the cold-activating mutations appear more likely to occur at conserved sites (Fig. 10). In the case of the three closely related homologs S41, S39, and SSII, all thermostabilizing mutations in S41 are located at positions that are variable, while all the cold-adaptive mutations in SSII are at positions that are either identical or highly conserved. Similarly, all but two of the thermostabilizing mutations in subtilisin E are located at variable sites, while three of the four cold-adaptive mutations in BPN′ are at conserved sites. The preference of thermostabilizing mutations for variable sites is also related to fact that most mutations were found on the protein surface, where sites are generally more variable.

Table VI lists proposed stabilization mechanisms for the mutations found during directed evolution on subtilisins E and S41, and *p*NB esterase. A variety of mechanisms are represented, including the introduction of salt bridges and hydrogen bonds, improved packing of nonpolar groups, helix stabilization, loop stabilization, and enhanced calcium binding. The diversity of stabilization strategies in Table VI is similar to that found in natural thermophilic proteins (see a list compiled in a recent review by Jaenicke and Bohm, 1998). As with studies of natural proteins, we see no single preferred mechanism or type of interaction for increasing protein stability. In the laboratory as well as in natural evolution, thermostabilization generally results from the combined effects of multiple, modestly stabilizing interactions. [Single substitutions that give dramatic increases in stability do exist (e.g., Williams *et al.*, 1999), but they are the exception rather than the rule.]

TABLE VI
Proposed Stabilization Mechanisms for Mutations in p*NB Esterase, Subtilisin E, and Subtilisin S41*

Mutation	Proposed mechanism of stabilization
pNB esterase	
A56V	Unknown
I60V	Loop stabilization, hydrogen bonding
T73K	Salt bridge
L144M	Loop stabilization
L313F	Aromatic-aromatic interaction
H322Y	Loop stabilization
A343V	Improved packing
M358V	Hydrophobic interactions
Y370F	Unknown
A400T	Unknown
G412E	Salt bridge
I437T	Hydrogen bonding, helix stabilization
T459S	Removal of exposed hydrophobic surface
Subtilisin E	
P14L	Increased helix propensity
N76D	Enhanced calcium binding
N118S	Hydrogen bonding, helix stabilization
S161C	Unknown
G166R	Hydrogen bonding
N181D	Hydrogen bonding, protein-solvent interactions
S194P	Hydrogen bonding, loop stabilization
N218S	Improved hydrogen bond, improved packing
Subtilisin S41	
S145I	Improved hydrophobic interactions
S175T	Possible loop-turn rigidification
K211P	Loop stabilization
R212A	Unknown
K221E	Unknown
N291I	Improved hydrophobic interactions
S295T	Unknown

It is similarly not possible to infer general rules of adaptation from the limited data currently available regarding the mutations responsible for cold adaptation in subtilisins SSII and BPN′. Both the K_M mutations and the k_{cat} mutations improve activity at low temperatures preferentially, but it is unclear how this is achieved. It was noted before that the magnitudes of the intermolecular forces that are responsible for binding change with temperature. It may be that the mutations affecting K_M have altered enzyme-substrate interactions to accommodate these changes.

However, a role for mobility cannot be ruled out. K_M mutations may facilitate molecular motions required for substrate recognition. Without a great deal more structural and dynamic data, we cannot identify with any certainty specific mechanisms by which the activity at low temperature increased.

In several of the directed evolution studies detailed above, enzymes were screened for activity against small synthetic compounds (s-AAPF-*p*NA for subtilisins E, S41, and SSII; *p*-nitrophenylacetate for *p*NB E) for which the enzyme active sites were never optimized during natural evolution. This raises the possibility that the increases in activity that were seen during directed evolution are artifacts resulting from the use of nonnatural substrates, and that simultaneous increases in both stability and activity would not be achieved during selection on more ''realistic'' substrates. The results for the evolution of thermostability in IPMDH argue against this, however, since in that case evolution was carried out on the natural substrate. It may be expected that screening on nonnatural substrates will influence the nature of the mutations that are discovered. In the case of SSII, none of the cold-activating mutations are present in the highly homologous psychrophiles S41 or S39, and this is possibly a consequence of screening on s-AAPF-*p*NA.

V. Stability, Flexibility, and Catalytic Activity

The inverse correlation between stability at high temperatures and activity at low temperatures is often explained by recourse to the concept of protein flexibility, the proposed link between the two properties. Directed evolution clearly demonstrates that the inverse correlation seen in natural proteins is not physically necessary, and that proteins possessing both high stability and efficient low-temperature activity can be readily obtained in the laboratory. What do these results say about the proposed relationship between stability, flexibility, and catalytic activity? That protein motions are important for both stability and activity is clear from years of theory and experiment. The question that remains to be answered is ''Which motions?'' Proteins exhibit motions on distance scales from 0.01 to 100 Å and time scales from 10^{-15} to 10^{3} seconds (Brooks *et al.,* 1988). These include local atomic fluctuations, side chain motions, loop rearrangements, rigid body motions between subunits or elements of secondary structure, and large-scale cooperative processes such as the folding-unfolding transition. Which of these motions are related to stability and which are related to catalytic activity, and how are they related to each other? We now consider these questions in greater detail, and suggest how directed evolution may help in providing answers.

A. *Protein Motions and Stability*

The earliest experimental results on mobility in proteins came from hydrogen exchange experiments. Similarly, the first evidence suggesting a link between protein flexibility and thermal stability came from hydrogen exchange studies on the small globular protein bovine pancreatic inhibitor (BPTI) (Wagner and Wuthrich, 1979). Comparing exchange rates for nine assigned protons between wild type and several chemically modified variants, it was found that variants with lower denaturation temperatures exhibited higher exchange rates. Later studies, such as that by Zavodzski *et al.* (1998) on 3-isopropylmalate dehydrogenase, have found a similar trend between exchange rates and thermal stability. Although these results are intriguing, their interpretation in terms of specific protein motions is difficult since the connection of hydrogen exchange to detailed protein dynamics is not yet clear.

It is thought that amide protons must be exposed to solvent in order to undergo exchange (Englander and Kallenbach, 1984), and the exchange behavior of buried protons is generally pictured in terms of some conformational rearrangement of the protein exposing a given proton to solvent followed by the actual exchange:

$$\textit{closed} \underset{k_{-1}}{\overset{k_1}{\Leftrightarrow}} \textit{open} \overset{k_2}{\Rightarrow} \textit{exchange.}$$

Under so-called EX2 conditions, in which the rate-limiting step is the actual exchange process, the observed exchange rate is proportional to the equilibrium constant (k_1/k_{-1}) for the protein ‘‘opening’’ event(s) that result in the exposure of a given proton. This equilibrium constant will represent the net effect of all protein motions that result in exposure of a particular amide proton. The exact nature of these motions remains somewhat controversial. Two pathways for exchange are generally proposed. In one, local conformational changes result in the exposure of a particular region, while the rest of the protein remains essentially in the native state. A second pathway involves exchange from the globally unfolded state. For a number of proteins there is evidence that one subset of residues exchanges by the first mechanism while another subset exchanges by the second, although in some cases both mechanisms may play a role (Key-Sun and Woodward, 1993; Bai *et al.*, 1995; Yamasaki *et al.*, 1998). Those residues that exchange primarily by the global unfolding mechanism should indeed be correlated with stability; in fact, their exchange rates can provide an estimate for the free energy of global unfolding (Mayo and Baldwin, 1993; Key-Sun and Woodward, 1993; Bai *et al.,* 1994; Huyghues-Despointes *et al.,* 1999). In a study of lysozyme, Delepierre *et al.* (1983) measured the exchange rates for the indole NH hydrogens of tryptophan residues as functions of temperature in the presence and absence of urea, for wild type and a chemically modified variant with increased thermostability. They found that the exchange behavior was best described by two processes: a low activation barrier process (presumably corresponding to out exchange from a near native state) that did not correlate with stability, and a high activation barrier process (corresponding to exchange from the globally unfolded state) that did. These proton-specific factors complicate the interpretation of H/D exchange experiments in which exchange rates are measured for the entire protein rather than specific protons.

In addition to these complications in interpreting H/D exchange data, it must be born in mind that hydrogen exchange provides a static measure of protein flexibility: proteins in solution exist as an ensemble of different conformations. The population of each conformation is determined by its Gibbs free energy according to the standard statistical thermodynamic relationship

$$P_i = \frac{\exp(-\Delta G_i/RT)}{Q}$$

where ΔG_i is the difference in Gibbs free energy between state i and a reference state, R is the gas constant, T is temperature, and Q is the partition function—the sum of the statistical weights of all accessible protein states. Under native conditions, the state of lowest free energy will generally be very close to the three-dimensional structure as determined by X-ray crystallography or NMR (this serves as the reference state), and this state will represent the most highly populated conformation. However, other states, particularly those close in free energy to the ground state, will also be populated. ''Flexibility,'' in the equilibrium sense of the word, is defined as the extent to which nonnative states are populated at equilibrium under a given set of conditions. It is this flexibility that is probed by hydrogen exchange experiments, since exchange will occur from those nonnative states that expose buried amide hydrogens (Hilser and Freire, 1996; Tang and Dill, 1998). There is not yet a clearly established connection between this equilibrium measure of flexibility and protein dynamics. When the average lengths and fluctuations of hydrogen bonds derived from a molecular dynamics simulation of BPTI were compared with the experimental exchange rates for all the amide protons in that protein, little or no correlation was found (Wagner and Wuthrich, 1982). More recently, a combined H/D exchange and ^{15}N relaxation study of ribonuclease HI found no correlation between the two measurements (Yamasaki *et al.,* 1998).

Other measures of protein flexibility have been found to correlate with thermal stability. One is resistance to proteolysis (Daniel *et al.,* 1982; Fontana, 1988). Another is the quenching of buried tryptophan fluorescence by acrylamide, used in a study by Varley and Pain (1991). Both these processes are mediated by the same combination of local and global unfolding events that determine rates of hydrogen exchange. Their rates will depend on the ability of another molecule, acrylamide or a proteolytic enzyme, to penetrate into normally buried regions of the protein in order to either quench fluorescence or cleave peptide bonds.

Nuclear magnetic relaxation rates measured by NMR spectroscopy provide another measure of protein mobility. Unlike H/D exchange, resistance to proteolysis, or tryptophan fluorescence quenching, the ^{15}N or ^{13}C relaxation rates are not determined solely by local or global unfolding, and so can probe smaller scale fluctuations in the native state. Further, nuclear relaxation measurements provide information on dynamics from the pico- to millisecond time scale. What is the relation of these motions to thermal stability? In an important recent study of wild-type RNAse H1 and a mutant whose melting temperature is 20.2°C higher, Yamasaki *et al.* (1998) found that ^{15}N relaxation rates for most residues were unaffected by the increase in stability. Furthermore, they identified amide protons that undergo relatively fast exchange in the wild-type enzyme—that is, those protons that are most likely to exchange through local fluctuations rather than global unfolding—and found that their H/D exchange rates were essentially the same in the wild-type and mutant enzymes. Thus they found no correlation between thermal stability and native state fluctuations, as measured by these two techniques.

Both the stability and the motions of proteins have been interpreted in terms of the ''energy landscape'' (Frauenfelder *et al.,* 1991; Onuchic *et al.,* 1997, Dill *et al.,* 1995; Abkevich *et al.,* 1994). The energy landscape is a high-dimensional hypersurface that maps the potential energy of a protein as a function of conformation. It is generally portrayed as being extremely rugged, with many peaks and valleys. The valleys correspond to local free energy minima or ''conformational substates,'' while the peaks correspond to the energetic barriers between these substates. Native state flexibility, then, is related to the distribution of conformational substates around the native state and the heights of the barriers between them. Thermodynamic stability, in contrast, is related to the free energy difference between the native state ensemble—which may include many conformational substates—and the denatured state ensemble (Fig. 20). From this point of view, there seems no reason to expect a correlation between stability and flexibility. This point was noted by Lazaridis *et al.* in a discussion of the dynamics and unfolding behavior of mesophilic and hyperthermophilic rubredoxins (Lazaridis *et al.,* 1998). These authors also point out that flexibility cannot, in itself, be offered as a reason for decreased stability since increased flexibility implies increased conformational entropy in the native state and should therefore be stabilizing. Increased flexibility may arise through a reduction of energetically favorable interactions such as van der Waals contacts, which will tend to decrease stability, but this is not necessarily always the case.

The results of directed evolution on subtilisin E, *p*NB E, and IPMDH suggest that, in analyzing thermal stability, it may be more useful to concentrate attention on specific motions that can act as initial events in unfolding, rather than on some general notion of flexibility. For both subtilisin E and *p*NB E, an important step in the evolution of thermal stability was the restriction of large-scale loop motions. Simulations of subtilisin E showed that, at high temperatures, a large-scale loop motion that exposed the core of the protein to solvent occurred, while a similar motion was lacking in the thermostabilized mutant (Colombo and Merz, 1999). This kind of loop motion that exposes the protein core has been identified in other protein systems as an initial event in unfolding (Lazaridis *et al.,* 1997; Caflisch and Karplus, 1995). Taken together, these studies suggest that fluctuations that expose the protein core to solvent

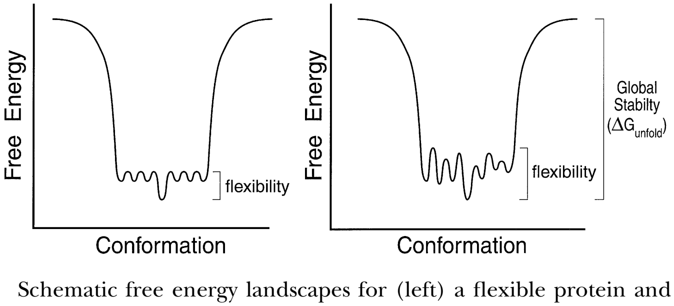


FIG. 20. Schematic free energy landscapes for (left) a flexible protein and (right) a more rigid protein. Flexibility is determined by the distribution of free energy minima and barriers around the native state, while global stability is a function of the free energy difference between the native state ensemble and the denatured state ensemble.

will be particularly important for thermostability. The H/D exchange of buried amide protons and the quenching of buried tryptophans by acrylamide are both, at least in part, mediated by exactly these kinds of fluctuations. It is not surprising, therefore, to find that they show a correlation with thermostability. On the other hand, flexibility as judged by the RMS deviations from the average structure is actually greater in the thermostable variant of subtilisin E and may in fact contribute to stabilization. Flexibility, as measured by ^{15}N relaxation, showed no correlation with stability in the case of RNase HI.

B. Protein Motions and Activity

It has been known for some time that molecular motions play an important role in substrate specificity and catalysis. The precise nature of these motions, however, remains unclear in most cases. Ever since Koshland put forth the ''induced fit'' model of enzyme-substrate recognition (Koshland, 1958), it has been appreciated that enzymes often must undergo conformational changes on binding their substrates. In certain cases in which structures of enzymes in their free and substrate-bound forms are available, specific conformational changes resulting from binding have been identified. Observed binding-induced conformational changes have recently been interpreted in statistical terms (Freire, 1998; Freire, 1999; Ma *et al.,* 1999). In this picture, the observed structural differences between the free and substrate-bound forms of an enzyme are not the product of a specific conformational change from one state to another induced by substrate binding. Rather, they are the result of a redistribution of the protein population among different conformational substates. The unbound structure represents the energetically most favorable conformation in the absence of substrate. However, as mentioned above, a range of nonnative conformations will also be populated under these conditions, and some of these nonnative conformations will be structurally similar to the protein's substrate-bound form. Since they are energetically less favorable than the native state, nonnative conformations will be populated to a much lower extent than the native state. The presence of substrate, however, will alter the balance of free energy between the various conformational substates. Those conformations resembling the substrate-bound form (''binding competent'' states) will now be energetically favored and thus become the most populated states. This perspective suggests a relationship between flexibility (in the equilibrium sense of the term) and function, at least as it refers to substrate binding. In order to effectively bind its substrate, an enzyme should be sufficiently ''flexible'' that under native conditions, binding-competent nonnative states will be populated to a reasonable degree. It is unclear at this point how general this picture of binding is. It has been suggested that the scheme described above will be most true for enzymes with broad substrate specificities, while the traditional lock-and-key picture will be more correct for highly specific enzymes (Ma *et al.,* 1999).

Far less is known about those enzyme motions required for catalysis as opposed to substrate recognition. Specific motions important to catalysis that have been identified in structural studies include loop movements which cover the active site in order to exclude solvent and stabilize intermediates (Kim *et al.,* 1997; Pompliano *et al.,* 1990) and rotations of subdomains with respect to each other (Sawaya and Kraut, 1997). In addition to specific conformational changes involving multiple protein domains, there is evidence that fluctuations, particularly near the active site, play a role. Mutations resulting in increased activity in hen eggwhite lysozyme were found to be associated with increased mobility of active site residues on the picosecond to millisecond time scales, although no specific conformational change was detected (Mine *et al.,* 1999). Several enzymes, including fructose-1,6-bisphosphatase, alkaline phosphatase, catalase, and DHFR are known to show increased activity in low concentrations of denaturants such as urea and GuHCl (Blakley, 1984; Zhao *et al.,* 1986; Duffy *et al.,* 1987; Tsou, 1998) In the case of DHFR, it has been demonstrated by limited proteolysis that the increased activity is accompanied by increased flexibility near the active site (Fan *et al.,* 1996).

The most direct evidence for a role for fluctuations in enzymatic catalysis comes from a recent study by Kohen *et al.* (1999) on alchohol dehydrogenase from the thermophilic bacterium *Bacillus stearothermophilus.* As with mesophilic ADH's (Cha *et al.,* 1989), quantum mechanical hydrogen tunneling was found to play a significant role in catalysis. According to models for tunneling through a rigid barrier, the contribution from tunneling should become increasingly significant with decreasing temperature. Experimentally, however, it was found that the contribution from tunneling decreased at and below 30°C (the organism's natural physiological temperature is 65°C). The authors interpreted this behavior in terms of decreased thermal fluctuations experienced by the enzyme at mesophilic temperatures: Some combination of protein motions that presumably facilitate tunneling occur with reduced frequency at room temperature and therefore the efficiency of tunneling decreases. As with most other enzymes, the precise nature of these thermally excited fluctuations that facilitate catalysis is unknown.

We have very briefly surveyed the types of protein motions thought to be associated with substrate recognition and catalysis. There is no *a priori* reason to expect that any of the structural fluctuations discussed should be connected with the kinds of large-scale unfolding events that are thought to be correlated with global stability. In terms of the energy landscape, adaptation for catalytic function should consist primarily of optimizing the distribution of conformational substates near the native state and the energetic barriers between them at a given temperature. This process might or might not be accompanied by changes in the free energy difference between the native state and denatured state. The studies on lysozyme and DHFR suggest that, whatever the role for thermal fluctuations in catalysis, it is primarily fluctuations near the active site that contribute. Indeed, it has been shown for lysozyme that stabilizing mutations located in the active site result in decreased catalytic activity (Shoichet *et al.,* 1995). Mobility at the active site could set a limit to an enzyme's stability if the active site is the weak point in the tertiary structure and thus serves as an initiation site for unfolding. However, there is not clear evidence that this is a general mechanism for protein unfolding. In the case of the small ribonuclease barnase, for example, molecular dynamics simulations of the denaturation process at 600°K (Caflisch and Karplus, 1994; Caflisch and Karplus, 1995) indicate that unfolding does not begin at the active site. Rather, during the first 30 picoseconds, the N-terminus and two loops (neither containing active site residues) begin to unfold, followed quickly by denaturation of the hydrophobic cores.

Results from the directed evolution of subtilisin E, *p*NB esterase, 3-isopropylmalate dehydrogenase, and subtilisin S41 demonstrate that it is possible to restrict those conformational motions that might lead to unfolding (such as the loop movement in subtilisin E) without interfering with the motions that are necessary for activity. As we have seen, the relationship between protein mobility, activity, and stability is still unclear. The dynamics of a mesophilic enzyme and its thermophilic homolog may differ in many ways and on many time scales. Some of these differences may be related to stability, some may be related to activity, and some may be related to neither. Just as proteins may have differences in their primary sequences that are functionally neutral, it is quite possible that they may exhibit differences in their dynamic properties that are unrelated to adaptation. Directed evolution can play an important role in identifying functionally important motions. As methods for studying biomolecular dynamics become increasingly powerful, it will be possible to characterize the diversity of structural fluctuations proteins experience. Directed evolution generates entire lineages of closely related variants, each showing an incremental change in stability and/or activity. These lineages contain information that is not available from studies of pairs of distantly related enzymes. Characterizing the dynamics of all members of such an artificial protein ''family'' may reveal that changes in mobility in some regions or on some time scales are strongly correlated with increases or decreases in activity/stability, while other changes are not. The principles of natural selection can thus be combined with biophysical techniques to pick out particular structural fluctuations from the otherwise overwhelming array that proteins typically experience. The phosphorescence studies of evolved thermostable *p*NB esterases (Section IV,C,1) represent a first step in this direction.

VI. Why Are Proteins Marginally Stable?

Researchers often point to the striking fact that natural proteins under near-physiological conditions are poised at the brink of instability (Privalov, 1979; Jaenicke and Bohm, 1998; Dill, 1990). For most proteins, the Gibbs free energy of stabilization at room temperature is between 5 and 20 kcal/mole (Dill, 1990), the equivalent of only a few weak intermolecular interactions (Jaenicke, 1990). From site-directed mutagenesis studies (Alber, 1989; Matthews, 1993) and directed evolution, we know that proteins can become more stable through point mutations. Why, then, should proteins be only marginally stable when higher stability is so easily obtained? The physical-chemical explanation is that higher levels of stability are not consistent with efficient catalysis. However, enzymes that are both highly active at low temperatures and highly stable can be generated when selective pressure is applied to both properties. Thus, we are again left with the question of why proteins generally exhibit such modest stability.

One explanation is immediately suggested by examining the effects of random mutations on protein stability. The results from screening random mutant libraries (such as those for *p*NB esterase and subtilisin S41) clearly show that the large majority of mutations are destabilizing. A similar conclusion was drawn previously by Sauer and co-workers, who investigated the tolerance of proteins for single and multiple random amino acid substitutions (Reidhaar-Olson and Sauer, 1988; Lim and Sauer, 1989; Bowie *et al.,* 1990; Lim and Sauer, 1991; Lim *et al.,* 1992; Gregoret and Sauer, 1998; Cordes and Sauer, 1999). When we recall that evolution involves the accumulation of multiple mutations over many generations, it becomes clear that, in the absence of constraints, protein stability is almost certain to decrease over time. We have seen examples of this in the laboratory, for example, during the evolution of subtilisin SSII (Section IV,D,1), where the net effect of multiple, cold-activating mutations was decreased stability.

In the absence of selection, the continued accumulation of mutations would eventually destroy the ability of a protein to fold at all. In nature, however, the biological requirement that proteins be functional (and therefore folded) will prevent proteins from reaching this point (Fig. 21). Viewed this way, it is more reasonable to assume that the marginal stability of natural proteins represents something close to the minimum that is consistent with biological function. This is not to say that natural proteins are literally minimally stable. It is generally possible to create destabilized mutants in the laboratory that still fold and function. However, a typical protein is only a few, modestly destabilizing mutations away from a point of ''minimal stability'' where one more minor decrease will compromise its ability to fold. Once a protein is close to its threshold stability, below which its ability to function may be compromised, its evolution is likely to proceed through the acceptance of modestly stabilizing and destabilizing mutations. Dramatically stabilizing mutations are rare and thus unlikely to occur. Dramatically destabilizing mutations will be eliminated by natural selection.

Support for this picture can be found in a study of the two closely related mesophilic proteins barnase and binase. These two extracellular ribonucleases are highly similar in structure, stability, and sequence, differing by only seventeen out of 110 amino acids. Serrano *et al.* (1993) introduced each mutation present in binase into barnase and measured

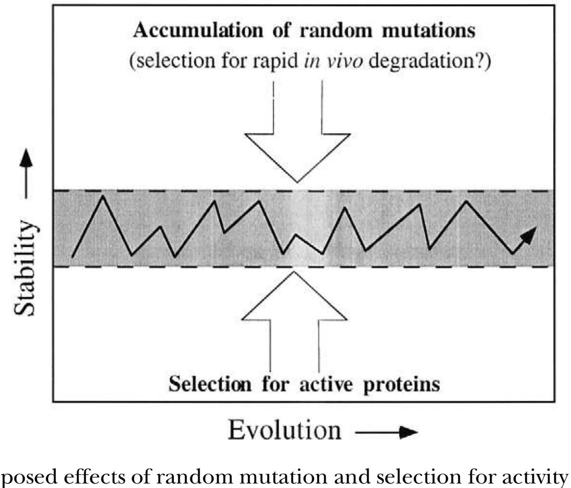

Accumulation of random mutations
(selection for rapid *in vivo* degradation?)
Stability
Selection for active proteins
Evolution

FIG. 21. Proposed effects of random mutation and selection for activity on the stability of an enzyme evolving at a given temperature. Stability fluctuates within a range determined by the destabilizing influence of accumulating random mutations and by the minimal stability that is required to remain folded and functional.

the changes in stability. They found that each mutation altered stability by no more than ± 1.1 kcal/mol, and that the effects of stabilizing and destabilizing mutations approximately cancelled each other. These results are consistent with the notion that, evolving at a given temperature, protein stability will fluctuate around a lower limit set by selection. Single, mildly destabilizing mutations will be accepted, but multiple uncompensated ones will not. Interestingly, one of the multiple mutants generated during this process was 3.3 kcal/mol more stable than wild type but retained full wild-type activity, suggesting, again, that stability need not come at the expense of catalytic activity.

Another possible explanation for the marginal stability of naturally occurring proteins is that excess stability is *disadvantageous* in the context of a living cell and has therefore been selected against (Benner and Ellington, 1990). An important aspect of cellular regulation, both in terms of the cell cycle and in terms of response to changes in the environment, is the regulation of the intracellular concentrations of enzymes. The rates of protein synthesis and degradation are balanced in order to maintain given enzyme concentrations, and can be altered to rapidly increase or decrease the amounts of key enzymes in the cell (Creighton, 1993). Highly stable enzymes that resist degradation could therefore be harmful to a cell by interfering with its ability to properly regulate enzyme concentrations. Of course, in light of what has been said regarding the effects of random mutations on protein stability, selection for instability in the cell may not be necessary. Because the work discussed here was performed mainly *in vitro,* it cannot distinguish between these possibilities.

VII. Why Are Thermophilic Enzymes Poorly Active at Low Temperature?

Having shown that high thermostability is not incompatible with high activity at low temperatures, we now consider alternative explanations for the poor low-temperature activity that is generally seen in natural thermophilic enzymes. Since the intrinsic rates of chemical reactions will be reduced at low temperatures, mesophilic and psychrophilic enzymes face a more difficult catalytic task than their thermophilic cousins. In fact, there is evidence that the effect of decreasing temperature on reaction rates is greater than generally believed. In discussing the influence of temperature on reaction rates, it is often stated that increasing the temperature from 20° to 30°C will approximately double the rate of a reaction (a ‘‘Q_{10}’’ value of 2) (Pauling, 1950). Recently, Wolfenden *et al.* (1999) determined the uncatalyzed rates for a number of reactions as functions of temperature. They found that most reactions showed Q_{10} values between 4 and 6, with some as high as 12. These strong temperature dependencies have significant implications for the relative proficiency required of enzymes operating at different temperatures. This can be seen by comparing the rates of hydrolysis of glycosides in the presence and absence of α-glucosidase (Wolfenden *et al.,* 1999). Decreasing the temperature from 60°C to 13°C increases the half-time of the uncatalyzed reaction from ~21,000 years to ~4,300,000 years, while the half-time for the catalyzed reaction increases from ~0.0007 seconds to ~0.01 seconds. To maintain a reaction rate at 13°C comparable to that observed at 60°C, an enzyme would require a rate enhancement of the uncatalyzed reaction of $\sim 1.3 \times 10^{17}$, compared to $\sim 9 \times 10^{14}$ for the enzyme at 60°C. This requirement for a three orders of magnitude greater rate enhancement at low temperature clearly places the catalytic machinery of psychrophilic enzymes under more severe selective pressure than thermophilic enzymes. In the absence of such pressure, there is no reason to expect that a thermophilic α-glucosidase would spontaneously evolve such a significant increase in rate enhancement.

A second factor to consider is the intrinsic increase in reaction rates caused by increasing temperature. An enzyme maintaining a rapid rate at low temperatures will exhibit extremely rapid reaction rates at elevated temperatures. Enzymes must function within the context of cellular metabolism, and such ultrafast enzymes may be disadvantageous.

VIII. Conclusions

To date, only a small number of directed evolution studies of temperature adaptation have been carried out. From this small data set we can nonetheless draw several interesting conclusions.

- When the enzyme is freed from the constraints of biological function, we can use laboratory evolution to explore the physically possible, and distinguish that from the biologically relevant. The apparent trade-off between thermal stability and low temperature activity is not physically necessary, at least for the properties and conditions investigated in the experiments. Enzymes displaying a high degree of stability at elevated temperatures and high activity at moderate or even low temperatures can be generated when selective pressure is applied to *both* properties.
- The lower stability of psychrophilic and mesophilic enzymes relative to their counterparts from thermophiles is likely to be the product of genetic drift, or have occurred during adaptation of other properties

such as catalytic activity. Because most mutations are deleterious, any property that is drifting will eventually drift down. Thus, thermal stability would be lost in the absence of selective pressure to maintain or increase it. Lacking detailed knowledge of the selective pressures experienced during evolution in nature, however, we cannot rule out the possibility that highly stable enzymes are actually disadvantageous in mesophilic and psychrophilic organisms and are thus subject to negative selection.

- Because we often do not know the selective pressures under which natural enzymes evolved, and because we cannot easily distinguish adaptive mutations from neutral mutations or mutations responding to other pressures, it will be very difficult to generate reliable design rules from sequence comparisons. For example, if we assume that early enzymes functioned at high temperatures, and stability was gradually lost as organisms colonized lower temperature environments, then sequence comparisons will yield little useful information for protein engineers who wish to improve the high-temperature performance of mesophilic enzymes. Laboratory-evolved enzymes, in contrast, offer a much cleaner system in which to study adaptive mechanisms. Implementing different evolutionary scenarios in the laboratory allows us to assess adaptive potential and identify specific adaptation mechanisms.
- Active and/or stable enzymes can be generated rapidly. Although natural proteins adapted to different temperatures typically differ from each other at many amino acid positions, only a small number of substitutions are required to confer high degrees of stability and/or low temperature activity. Some parallel nature, some do not.

Directed evolution as a tool to probe the basis of protein structure, stability, and function is in its infancy, and many fruitful avenues of research remain to be explored. Studies so far have focused on proteins that unfold irreversibly, making detailed thermodynamic analysis impossible. The application of these methods to reversibly folding proteins could provide a wealth of information on the thermodynamic basis of high temperature stability. A small number of studies on natural thermophilic proteins have identified various thermodynamic strategies for stabilization. Laboratory evolution makes it possible to ask, for example, whether proteins have adopted these different strategies by chance, or whether certain protein architectures favor specific thermodynamic mechanisms. It will also be possible to determine how other selective pressures, such as the requirement for efficient low temperature activity, influence stabilization mechanisms. The combination of directed evolution with high resolution structural studies and detailed characterization of dynamics promises to provide insights into the molecular basis of stability and catalysis. The work reviewed here represents only a first step toward this goal.

Although proteins in themselves are important resources for biotechnology and fascinating objects for physiochemical studies, their true importance stems from their role as functional components of living organisms. Here, we have stressed the potential of directed evolution to adapt proteins for function under nonbiological conditions. A goal for the future, however, is to determine how specific properties measured *in vitro* affect biological function, and the fitness of the organism, by introducing laboratory-evolved proteins back into their host organisms. Such studies may help, for example, define the role of negative selection. For example, by introducing thermostabilized enzymes that nonetheless retain wild-type activity at moderate temperatures into a mesophilic host, it should be possible to test whether excessive stability in enzymes is detrimental to the organism's survival.

Directed evolution can also be used to probe the boundaries of protein function, for example, the role of protein stability in setting the upper temperature limits for life. At present, the most thermophilic organisms known grow up to temperatures of ~113°C. The true maximum temperature for life, whatever it turns out to be, may be set by any one or combination of factors, including the stability of lipid bilayers, RNA or DNA, small molecules such as ATP, or proteins and enzymes. Although polypeptides are known to undergo degradation reactions near 100°C, their folded conformations provide substantial protection against these reactions (Daniel, 1996). The true upper limit for protein stability and function is therefore still an open question and may never even have been explored during natural evolution. It may therefore be possible to evolve enzymes that are stable and active at temperatures above those at which even the most hyperthermophilic microorganisms can survive.

References

Abkevich, V. I., Gutin, A. M., and Shakhnovich, E. I. (1994). *J. Chem. Phys.* **101,** 6052–6062.

Adams, M., and Kelly, R. (1998). *TIBTECH* **16,** 329–332.

Aghajari, N., Feller, G., Gerday, C., and Haser, R. (1998). *Protein Sci.* **7,** 564–572.

Akanuma, S., Yamagishi, A., Tanaka, N., and Oshima, T. (1998). *Protein Sci.* **7,** 678–705.

Alber, T. (1989). *Ann. Rev. Biochem.* **58,** 765–798.

Aoshima, M., and Oshima, T. (1997). *Protein Eng.* **10,** 249–254.

Arnold, F. H. (1998). *Nature Biotechnology* **16,** 617–618.

Arnold, F. H., and Wintrode, P. L. (1999). In *Encyclopedia of Bioprocess Technology: Fermentation, Biocatalysis, and Bioseperation* (Flickinger, M. C., and Drew, S. W. eds.), pp. 971–987. John Wiley & Sons, New York.

Auerbach, G., Huber, R., Grattinger, M., Zaiss, K., Schurig, H., and Jaenicke, R. (1997a). *Structure* **5,** 1475–1483.

Auerbach, G., Jacob, U., Grattinger, M., Zaiss, K., Schurig, H., and Jaenicke, R. (1997b). *J. Biol. Chem.* **378,** 327–329.

Auerbach, G., Ostendorp, R., Prade, L., Korndorfer, I., Huber, R., and Jaenicke, R. (1998). *Structure* **6,** 769–781.

Bai, Y. W., Milne, J. S., Mayne, L., and Englander, S. W. (1994). *Proteins* **20,** 4–14.

Bai, Y., Sosnick, T. R., Mayne, L., and Englander, W. S. (1995). *Science* **269,** 192–197.

Beadle, B. M., Baasw, W. A., Wilson, D. B., Gilkes, N. R., and Shoichet, B. K. (1999). *Biochemistry* **38,** 2570–2576.

Benner, S. A. (1989). *Chem. Rev.* **89,** 789–806.

Benner, S. A., and Ellington, A. D. (1990). *Crit. Rev. Biochem. Mol.* **23,** 369–426.

Blakley, R. L. (1984). In *Folates and Pterins* (Blakley, R. L., and Benkovic, S. J., eds.) Vol. 1, pp. 191–253. John Wiley & Sons, New York.

Bowie, J. U., Reidhaar-Olson, J. F., Lim, W. S., and Sauer, R. T. (1990). *Science* **247,** 1306–1310.

Braxton, S., and Wells, J. (1992). *Biochemistry* **31,** 7796–7801.

Brooks, C. L., Karplus, M., and Pettit, B. M. (1988). *Adv. Chem. Phys.* **71,** 1–259.

Cadwell, W. P. C., and Joyce, G. (1994). *PCR Methods Appl.* **3,** S136–S140.

Caflisch, A., and Karplus, M. (1994). *Proc. Natl. Acad. Sci. USA* **91,** 1746–1750.

Caflisch, A., and Karplus, M. (1995). *J. Mol. Biol.* **252,** 672–708.

Cha, Y., Murray, C. J., and Klinman, J. P. (1989). *Science* **243,** 1325–1330.

Chen, K., and Arnold, F. H. (1993). *Proc. Natl. Acad. Sci. USA* **90,** 5618–5622.

Colombo, G., and Merz, K. M. (1999). *J. Am. Chem. Soc.* **121,** 6895–6903.

Cordes, M. H. J., and Sauer, R. T. (1999). *Protein Sci.* **8,** 318–325.

Creighton, T. E. (1993). *Proteins: Structures and Molecular Properties.* W.H. Freeman & Co., New York.

Dahiyat, B. I., and Mayo, S. L. (1997). *Proc. Natl. Acad. Sci. USA* **94,** 10172–10177.

Dams, T., and Jaenicke, R. (1999). *Biochemistry* **38,** 9169–9178.

Daniel, R. M., Cowan, D. A., Morgan, H. W., and Curran, M. P. (1982). *Biochem. J.* **207,** 641–644.

Daniel, R. M. (1996). *Enzyme Microb. Technol.* **19,** 74–79.

Daniel, R. M., and Danson, M. J. *Methods in Enzymology.* In press.

Davail, S., Feller, G., Narinx, E., and Gerday, C. (1994). *J. Biol. Chem.* **269,** 17448–17453.

Dayhoff, M. (1978). Atlas of Protein Sequence and Structure. Vol. 5, Supplement 3. National Biomedical Research Foundation. Washington, D.C.

Delepierre, M., Dobson, C. M., Selvarajah, S., Weden, R. E., and Poulsen, F. M. (1983). *J. Mol. Biol.* **168,** 687–692.

Dill, K. A. (1990). *Biochemistry* **29,** 7133–7155.

Dill, K. A., Bromberg, S., Yue, K. Z., Fiebig, K. M., Yee, D. P., Thomas, P. D., and Chan, H. S. (1995). *Protein Sci.* **4,** 561–602.

Dobzhansky, T. (1973). *American Biology Teacher* **35,** 125–129.

Doster, W., Cusack, S., and Petry, W. (1989). *Nature* **337,** 754–756.

Duffy, T. H., Beckman, S. B., Peterson, S. M., Vitols, K. S., and Huennekens, M. (1987). *J. Biol. Chem.* **262,** 7028–7033.

Elcock, A. H. (1998). *J. Mol. Biol.* **284,** 489–502.

Englander, S. W., and Kallenbach, N. R. (1984). *Quart. Rev. Biophys.* **16,** 521–655.

Fan, Y. X., Ju, M., Zhou, J. M., and Tsou, C. L. (1996). *Biochem. J.* **315,** 97–102.

Feller, G., and Gerday, C. (1997). *Cell. Mol. Life Sci.* **53,** 830–841.

Feller, G., d'Amico, D., and Gerday, C. (1999). *Biochemistry* **38,** 4613–4619.

Fontana, A. (1988). *Biophys. Chem.* **29,** 181–193.

Fontana, A., De Filippis, V., Polverino de Laurento, P., Scaramella, E., and Zambonon, M. (1998). In *Stability and Stabilization of Biocatalysts* (Ballestros, A., Plou, F. J., Iborra, J. L., and Halling, P. J., eds.), pp. 277–294. Elsevier Science B. V, Amsterdam.

Frauenfelder, H., Sligar, S. G., and Wolynes, P. G. (1991). *Science* **254,** 1598–1603.

Freire, E. (1998). *Adv. Prot. Chem.* **51,** 255–279.

Freire, E. (1999). *Proc. Natl. Acad. Sci. USA* **96,** 10118–10122.

Gerday, C., Aittaleb, M., Arpigny, J. L., Baise, E., Chessa, J. P., Garsoux, G., Petrescu, I., and Feller, G. (1997). *Biochem. Biophys. Acta* **1342,** 119–131.

Gershenson, A., Schauerte, J. A., Giver, L., and Arnold, F. H. (2000). *Biochemistry* **39,** 4658–4665.

Gershenson, A., and Arnold, F. H. in *Genetic Engineering, Principles and Methods.* In press.

Giver, L., Gershenson, A., Freskagard, P. O., and Arnold, F. H. (1998). *Proc. Natl. Acad. Sci. USA* **95,** 12809–12813.

Gregoret, L. M., and Sauer, R. T. (1998). *Folding & Design* **3,** 119–126.

Hatanaka, H., Tanimura, R., Katoh, S., and Inagaki, F. (1997). *J. Mol. Biol.* **268,** 922–933.

Hendsch, Z. S., and Tidor, B. (1994). *Protein Sci.* **3,** 211–226.

Hennig, M., Darimont, B., Sterner, R., Kirschner, K., and Jansonius, J. N. (1995). *Structure* **3,** 1295–1306.

Hensel, R., Jakob, I., Scheer, H., and Lottspeich, R. (1992). *Biochem. Soc. Symp.* **58,** 127–133.

Hilser, V. J., and Freire, E. (1996). *J. Mol. Biol.* **262,** 756–772.

Hinz, H. J., and Jaenicke, R. (1975). *Biochemistry* **14,** 24–27.

Hochachka, P. W., and Somero, G. N. (1984). *Biochemical Adaptation.* Princeton University Press, Princeton NJ.

Hollien, J., and Marqusse, S. (1999). *Proc. Natl. Acad. Sci. USA* **96,** 13674–13678.

Huyghues-Despointes, B. M. P., Scholtz, J. M., and Pace, C. N. (1999). *Nat. Struct. Biol.* **6,** 910–912.

Iwasaki, Y. H., Numata, K., Yamagishi, A., Yutani, K., Sakurai, M., Tanaka, N., and Oshima, T. (1996). *Protein Sci.* **5,** 511–516.

Jaenicke, R. (1990). *Phil. Trans. R. Soc. Lond.* B **326,** 535–553.

Jaenicke, R., and Bohm, G. (1998). *Curr. Opin. Struct. Biol.* **8,** 738–748.

Kast, P., and Hilvert, D. (1997). *Curr. Opin. Struct. Biol.* **7,** 470–479.

Kauzmann, W. (1959). *Adv. Prot. Chem.* **14,** 1–63.

Key-Sun, K., and Woodward, C. (1993). *Biochemistry* **32,** 9606–9613.

Kim, S. W., Cha, S. S., Cho, H. S., Kim, J. S., Ha, H. C., Cho, M. J., Joo, S., Kim, K. K., Choi, K. Y., and Oh, B. I. T. (1997). *Biochemistry* **36,** 14030–14036.

Kim, S. Y., Hwang, K. Y., Kim, S. H., Sung, H. C., Han, Y. S., and Cho, Y. (1999). *J. Biol. Chem.* **274,** 11761–11767.

Kimura, M. (1968). *Nature* **217,** 624–626.

Kimura, I. (1983). *The Neutral Theory of Molecular Evolution.* Cambridge University Press, Cambridge.

Kohen, A., Cannio, R., Bartolucci, S., and Klinman, J. P. (1999). *Nature* **399,** 496–499.

Korndorfer, I., Steipe, B., Huber, R., Tomschy, A., and Jaenicke, R. (1995). *J. Mol. Biol.* **246,** 512–521.

Koshland, D. E. (1958). *Proc. Natl. Acad. Sci. USA* **44,** 98–99.

Kraulis, P. J. (1991). *J. Appl. Crystallogr* **24,** 946–950.

Lazaridis, T., Lee, I., and Karplus, M. (1997). *Protein Sci.* **6,** 2589–2605.

Leung, D., Chen, E., and Goeddel, D. (1989). *Technique* **1,** 11–15.

Liao, H., Mckenzie, T., and Hageman, R. (1986). *Proc. Natl. Acad. Sci. USA* **83,** 576–580.

Lim, W. A., and Sauer, R. T. (1989). *Nature* **339,** 31–36.

Lim, W. A., and Sauer, R. T. (1991). *J. Mol. Biol.* **219,** 359–376.

Lim, W. A., Farruggio, D. C., and Sauer, R. T. (1992). *Biochemistry* **31,** 4324–4333.

Low, P. S., Bada, J. L., and Somero, G. N. (1973). *Proc. Natl. Acad. Sci. USA* **70,** 430–432.

Ma, B., Kumar, S., Tsai, C. J., and Nussinov, R. (1999). *Prot. Eng.* **12,** 713–720.

Macedo-Ribeiro, S., Darimont, B., Sterner, R., and Huber, R. (1996). *Structure* **4,** 1291–1301.

Makhatadze, G. I., and Privalov, P. L. (1995). *Adv. Prot. Chem.* **47,** 307–425.

Martinez, M. A., Pezo, V., Marliere, P., and Wain-Hobson, S. (1996). *EMBO J.* **15,** 1203–1210.

Matsumura, M., and Aiba, S. (1985). *J. Biol. Chem.* **260,** 15298–15303.

Matthews, B. W. (1993). *Ann. Rev. Biochem.* **62,** 139–160.

Mayo, S. L., and Baldwin, R. L. (1993). *Science* **262,** 873–876.

Mine, S., Tate, S., Ueda, T., Kainosho, M., and Imoto, T. (1999). *J. Mol. Biol.* **286,** 1547–1565.

Miyazaki, K., and Arnold, F. H. (1999). *J. Mol. Evol.* **49,** 716–720.

Miyazaki, K., Wintrode, P. L., Grayling, R., and Arnold, F. H. (2000). *J. Mol. Biol.* **297,** 1015–1026.

Moore, J. C., and Arnold, F. H. (1996). *Nat. Biotech.* **14,** 458–467.

Morawski, B., Lin, Z., Cirino, P., Joo, H., Bandara, G., and Arnold, F. H. (2000). *Protein Eng.* **13,** 377–384.

Narinx, E., Baise, E., and Gerday, C. (1997). *Protein Eng.* **10,** 1271–1279.

Onuchic, J. N., Lutheyschulten, Z., and Wolynes, P. G. (1997). *Annu. Rev. Phys. Chem.* **48,** 545–600.

Oshima, T. (1994). *Curr. Opin. Struct. Biol.* **4,** 623–628.

Pace, C. N., Bret, S. A., Mcnutt, M., and Gajiwala, K. (1996). *FASEB J.* **10,** 75–83.

Pauling, L. *College Chemistry,* pp. 410. Freeman, New York.

Pfeil, W., Gesierich, U., Kleemann, G. R., and Sterner, R. (1997). *J. Mol. Biol.* **272,** 591–596.

Pompliano, D. L., Peyman, A., and Knowles, J. R. (1990). *Biochemistry* **29,** 3186–3194.

Pontiliano, M, Whitlow, M., Wood, J. F., Dodd, S. W., Hardmann, K. D., Rollence, M. L., and Btyan, P. (1989). *Biochemistry* **28,** 7205–7213.

Privalov, P. L. (1979). *Adv. Prot. Chem.* **33,** 167–241.

Privalov, P. L., and Gill, S. J. (1988). *Adv. Prot. Chem.* **39,** 191–234.

Querol, E., Perez-Pons, J. A., and Mozo-Villarias, A. (1996). *Protein Eng.* **9,** 265–271.

Rasmussen, B. F., Stock, A. M., Ringe, D., and Petsko, G. A. (1992). *Nature* **357,** 423–424.

Reidhaar-Olson, J., and Sauer, R. T. (1988). *Science* **241,** 53–57.

Russel, R. J., and Taylor, G. L. (1995). *Curr. Op. Biotechnol.* **6,** 370–374.

Russell, R. J. M., Gerike, U., Danson, M. J., Hough, D. W., and Taylor, G. (1998). *Structure* **6,** 351–361.

Sawaya, M. R., and Kraut, J. (1997). *Biochemistry* **36,** 586–603.

Schmid, F. X., Hinz, H. J., and Jaenicke. R. (1976). *Biochemistry* **15,** 3052–3059.

Serrano, L., Day, A. G., and Fersht, A. R. (1993). *J. Mol. Biol.* **233,** 305–312.

Shakhnovich, E. I., and Finkelstein, A. V. (1989). *Biopolymers* **28,** 1667–1680.

Shoichet, B. K., Baase, W. A., Kuroki, R., and Matthews, B. W. (1995). *Proc. Natl. Acad. Sci. USA* **92,** 452–456.

Siezen, R. J., and Leunissen, J. A. M. (1997). *Protein Sci.* **6,** 501–523.

Skandalis, A., Encell, L. P., and Loeb, L. A. (1997). *Chem. & Biol.* **4,** 889–898.

Somero, G. N. (1995). *Annu. Rev. Physiol.* **57,** 43–68.

Spiller, B., Gershenson, A., Arnold, F. H., and Stevens, R. C. (1999). *Proc. Natl. Acad. Sci. USA* **96,** 12305–12310.

Stemmer, W. P. (1994a). *Nature* **370,** 389–391.

Stemmer, W. P. (1994b). *Proc. Natl. Acad. Sci. USA* **91,** 10747–10751.

Taguchi, S., Ozaki, A., and Momose, H. (1998). *Appl. Environ. Microbiol.* **64,** 492–495.

Taguchi, S., Ozaki, A., Nonaka, T., Mitsui, Y., and Momose, H. (1999). *J. Biochem.* (*Tokyo*) **126,** 689–693.

Tang, K. E. S., and Dill, K. A. (1998). *J. Biomol. Struct. Dyn.* **16,** 397–411.

Teplyakov, A. V., Kuranova, I. P., Harutyunyan, E. H., Vainshtein, B. K., Frommel, C., Hohne, W. E., and Wilson, K. S. (1990). *J. Mol. Biol.* **214,** 261–279.

Thompson, M. J., and Eisenberg, D. (1999). *J. Mol. Biol.* **290,** 595–604.

Tilton, R. F., Dewan, J. C., and Petsko, G. A. (1992). *Biochemistry* **31,** 2469–2481.

Tsou, C. (1998). *Biochemistry* (*Moscow*) **63,** 253–258.

Varley, P. G., and Pain, R. H. (1991). *J. Mol. Biol.* **220,** 531–538.

Vielle, C., and Zeikus, J. (1996). *Trends Biotechnol.* **14,** 183–191.

Vogt, G., and Argos, P. (1997). *Folding & Design* **2,** S40–S46.

Voordouw, G., Milo, C., and Roche, R. S. (1976). *Biochemistry* **15,** 3716–3724.

Wagner, G., and Wuthrich, K. (1979). *J. Mol. Biol.* **130,** 31–37.

Wagner, G., and Wuthrich, K. (1982). *J. Mol. Biol.* **160,** 343–361.

Waldburger, C. D., Schildbach, J. F., and Sauer, R. T. (1995). *Nature Struct. Biol.* **2,** 122–128.

Wallon, G., Kryger, G., Lovett, S. T., Oshima, T., Ringe, D., and Petsko, G. A. (1997). *J. Mol. Biol.* **266,** 1016–1031.

Watanabe, K., Masuda, T., Ohashi, H., Mihara, H., and Suzuki, Y. (1994). *Eur. J. Biochem.* **226,** 277–283.

Wati, M. R., Thanabalu, T., and Porter, A. G. (1997). *Biochem. Biophys. Acta* **1352,** 56–62.

Williams, J. C., Zeelen, J. P., Neubauer, G., Vriend, G., Backmann, J., Michels, P. A. M., Lambeir, A. M., and Wierenga, R. K. (1999). *Protein Eng.* **12,** 243–250.

Wintrode, P. L., Miyazaki, K., and Arnold, F. H. Submitted.

Wolfenden, R., Snider, M., Ridgway, C., and Miller, B. (1999). *J. Am. Chem. Soc.* **121,** 7419–7420.

Yamasaki, K., Furukawa, A. A., and Kanaya, S. (1998). *J. Mol. Biol.* **277,** 707–722.

Yip, K. S. P., Stillman, T. J., Britton, K. L., Artymiuk, P. J., Baker, P. J., Sedelnikova, S. E., Engle, P. C., Pasquo, A., Chiaraluce, R., and Consalvi, V. (1995). *Structure* **3,** 1147–1158.

Yip, K. S. P., Britton, K. L., Stillman, T. J., Lebbink, J., de Vos, W. M., Robb, F. T., Vetriani, C., Maeder, D., and Rice, D. W. (1998). *Eur. J. Biol.* **255,** 336–346.

Zaiss, K., Schurig, H., and Jaenicke, R. (1998). In *Biocalorimetry: Application of Calorimetry in Biological Sciences* (Ladbury, J., and Chowdhry, B. Z. Sussex, eds.), pp. 283–293. John Wiley & Sons, New York.

Zavodszky, P., Kardos, J., Svingor, A., and Petsko, G. A. (1998). *Proc. Natl. Acad. Sci. USA* **95,** 7406–7411.

Zhao, F. K., Shi, J. P., and Xu, G. J. (1986). *Sci. Sin.* **28B,** 599–607.

Zhao, H., and Arnold, F. H. (1997a). *Proc. Natl. Acad. Sci. USA* **94,** 7997–8000.

Zhao, H., and Arnold, F. H. (1997b). *Nucleic Acids Research* **25,** 1307–1308.

Zhao, H., and Arnold, F. H. (1999). *Protein Eng.* **12,** 47–53.

Zhao, H., Giver, L., Shao, Z., Affholter, J. A., and Arnold, F. H. (1998). *Nature Biotechnol.* **14,** 258–261.

This Page Intentionally Left Blank

STRUCTURAL ANALYSIS OF AFFINITY MATURED ANTIBODIES AND LABORATORY-EVOLVED ENZYMES

By M. CECILIA ORENCIA, MICHAEL A. HANSON, and RAYMOND C. STEVENS

Departments of Molecular Biology and Chemistry, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037

I. Introduction

The rational design of protein structure and function in the laboratory has not been overly successful, although various degrees of progress are being observed in many laboratories. In contrast, the processes of immune system affinity maturation and directed evolution have been successful in making very minor modifications to protein structure, with drastically enhanced physiochemical effects. In order to better understand why the more ''irrational'' design approach used by the immune system and directed evolution experiments are so successful, we have analyzed a series of experiments that determine the ''before'' and ''after'' three-dimensional protein structures. Ideally, we hope this analysis will provide insight into how the immune system and directed evolution accomplish successful protein engineering. To date, we know that mutations away from the active site have very pronounced effects in protein function, with very minor structural changes, especially to the active site cavity shape. It is surprising that most of the evolved proteins that have been studied so far have a very low number of mutations in the active site residues themselves. These global mutational changes are

0065-3233/00 $35.00

currently very difficult to predict *a priori,* and many lessons can be learned from the structural analysis of affinity matured and directed evolution derived enzymes.

II. Structural Studies of Antibody Affinity Maturation

The characteristic combinatorial assembly of gene segments, followed by somatic mutations generated during affinity maturation, make the immune system one of the most effective diversity-generating mechanisms known in nature. It is the exploitation of this seemingly boundless potential for diversity that characterizes the field of catalytic antibodies. The enormous diversity possible in the immune system derives from the shuffling of gene segments, which recombine to form the functional antibody, a di-chain molecule consisting of light (L) and heavy (H) chains (Fig. 1). The process of somatic mutation in affinity maturation

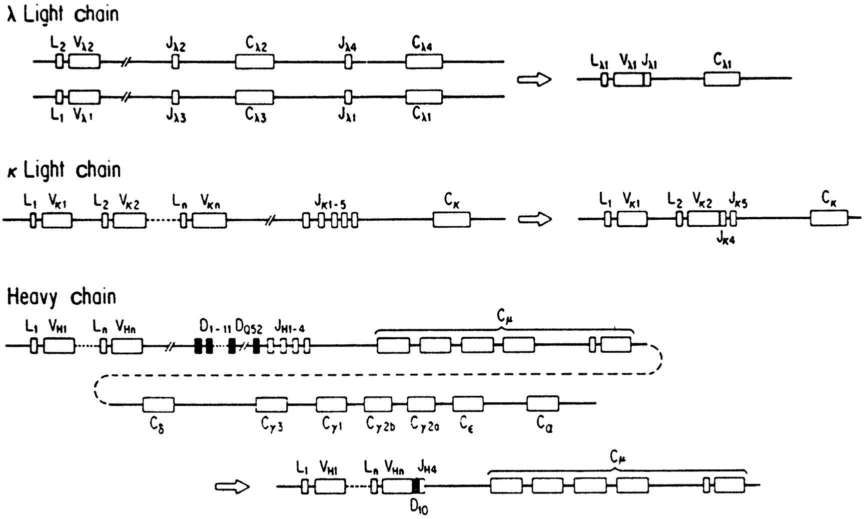


Fig. 1. The arrangement of germline genes in mice (from Tonegawa, 1983). The exact number of variable heavy genes is not known, but is somewhere between 200 and 1000. This results in between 12,000 and 60,000 different possible heavy chains, leading to possible library sizes on the order of 10^7–10^8 (data are for mice and are taken from Seidman and Leder, 1978a; Seidman *et al.* 1978b). Somatic mutation and junctional diversity further augment the size of this theoretical library to sizes that far exceed the number of B cells in a mouse (Rajewsky, 1996). For example, if any ten residues are mutated to any of the other nineteen amino acids, library size is increased by a factor of 19^{10}. Thus, diversity in the immune system is not limited by diversity-generating mechanisms.

can achieve enhanced specificity toward reaction transition state analogs and increased activity for a selected type of reaction.

Of the more than 100 catalytic antibodies known to catalyze chemical reactions, eighteen different antibodies have been structurally characterized, covering nine different chemical reactions (Table I). These include the ferrochelatase (Romesberg *et al.,* 1998a), Diels-Alderase (Heine *et al.,* 1998; Romesberg *et al.,* 1998b), cyclase (Lesberg *et al.,* 1998; Paschall *et al.,* 1999), aldolase (Barbas *et al.,* 1997), endo-tet cyclization (Gruber *et al.,* 1999), chorismate mutase (Haynes *et al.,* 1994a; Haynes *et al.,* 1994b; Ulrich *et al.,* 1997; Ulrich and Schultz, 1998; Mundorff *et al.,* 2000), sulfide-oxidase (Hsieh-Wilson *et al.,* 1996), and esterase catalytic antibodies (Golinelli-Pimpaneau *et al.,* 1994; Charbonnier *et al.,* 1995; Zhou *et al.,* 1994; Patten *et al.,* 1996; Wedemayer *et al.,* 1997a; Wedemayer *et al.,* 1997b; Buchbinder *et al.,* 1998; Thayer *et al.,* 1999; Kristensen *et al.,* 1998; Gigant *et al.,* 1997; Charbonnier *et al.,* 1997). Of these eighteen different antibodies, four have been structurally studied at both mature and germline levels.

A. *Esterolytic (48G7) Antibody Maturation*

Over half of the structurally characterized catalytic antibodies are from the esterolytic class of reactivity. This discussion will focus on 48G7, the most exemplary antibody from this class. In this antibody, the bound and free forms of both germline and mature antibodies have been structurally characterized, allowing for the first structural analysis of the affinity maturation process in a catalytic antibody system (Patten *et al.,* 1996; Wedemayer *et al.,* 1997a; Wedemayer *et al.,* 1997b).

Antibody 48G7, an esterolytic antibody raised against a *p*-nitrophenyl phosphonate transition-state analog, undergoes nine somatic mutations during the affinity maturation process (Patten *et al.,* 1996). Notably, none of these mutations contact the hapten, but they do elicit a 30,000-fold increase in K_m (135 to 0.0045 μM) (Patten *et al.,* 1996). Two points from this study are of special interest. First, the change in affinity is shown to be off-rate dependent, which is in direct conflict with kinetic theories of somatic maturation that support the biologically reasonable notion that competition for antigen drives selection for fast on rates (Foote and Milstein, 1991). Second, it reinforces the notion that somatic mutations distant from the combining site have a profound effect on specificity and activity.

Site-directed mutagenesis of 48G7 does not identify a small subset of mutations that are sufficient to induce mature level binding when introduced into the germline antibody (Patten *et al.,* 1996; Wedemayer

TABLE I

Structurally Characterized Catalytic Antibodies[a]

Antibody	Reaction type	PDB numbers	Free/liganded	Resolution (Å)	Reference
48G7	esterolytic	1GAF (affinity matured)	hapten-bound	1.95	Patten *et al.,* 1996
		1HKL (affinity matured)	apo	2.7	Wedemayer *et al.,* 1997a Wedemayer *et al.,* 1997b
		1AJ7 (germline)	hapten-bound	2.1	
		2RCS (germline)	apo	2.1	
17E8	esterolytic	1EAP	hapten-bound	2.5	Zhou *et al.,* 1994
CNJ206	esterolytic	1KNO	hapten-bound	3.2	Golinelli-Pimpaneau *et al.,* 1994
		2GFB	apo	3.0	Charbonnier *et al.,* 1995
29G11	esterolytic	1A0Q	hapten-bound	2.3	Buchbinder *et al.,* 1998
43C9 -scFv	aryl amidolytic, esterolytic	43C9	apo	2.2	Thayer *et al.,* 1999
43C9 -scFv	aryl, amidolytic esterolytic	43CA	*p*-nitrophenol bound	2.3	Thayer *et al.,* 1999
6D9	esterolytic	1HYX	TS analog-bound form I	1.8	Kristensen *et al.,* 1998
		1HYY	TS analog-bound form II	1.8	
D2.3 D2.4 D2.5	esterolytic	1YEC (D2.3)	[PNB] TS analog-bound	1.9	Gigant *et al.,* 1997
		1YEF (D2.3)	[PNC] substrate-analogue bound	2.0	Charbonnier *et al.,* 1997
		1YEG (D2.3)	[BPN] product-bound	2.0	
		1YEH (D2.3)	apo	2.55	
		1YEI (D2.3)	PNP-bound	1.9	
		1YEK (D2.3)	PNP-bound	2.1	
		1YED (D2.4)	[PNB] TS analog-bound	3.1	
		1YEE (D2.5)	[PNB] TS analog-bound	2.2	

7G12	metal chelatase	3FCT	hapten-bound	2.4	Romesberg *et al.,* 1998a
33F12	aldolase	1AXT	apo	2.15	Barbas *et al.,* 1997
HA5-19A4	cyclase	1CF8	hapten-bound	2.7	Lesberg *et al.,* 1998 Paschall *et al.,* 1999
1F7	chorismate mutase (Claisen rearrangement)	1FIG	hapten-bound	3.0	Haynes *et al.,* 1994a Haynes *et al.,* 1994b
AZ-28	chorismate mutase (oxy-Cope rearrangement)	1A (mature)	hapten-bound	2.6	Ulrich *et al.,* 1997
		submitted (mature)	apo	2.2	Ulrich and Schultz, 1998
		submitted (germline)	hapten-bound	2.0	
		submitted (germline)	apo	2.0	
28B4	sulfide-oxidase	1KEL	hapten-bound	1.9	Hsieh-Wilson *et al.,* 1996
		1KEM	apo	2.2	
5C8	disfavored endo-tet cyclizations	15C8	apo	2.5	Gruber *et al.,* 1999
		25C8	hapten-bound	2.0	
		35C8	inhibitor-bound	2.0	
39A11	Diels-Alderase	1A4K (mature)	hapten-bound	2.4	Romesberg *et al.,* 1998b
		1A4J (germline)	apo	2.1	
13G5	Diels-Alderase	1A3L	inhibitor-bound	1.95	Heine *et al.,* 1998

[a] Eight classes of antibody catalysis have been structurally characterized, several more than once [esterase (7), chorismate mutase (2), Diels-Alder(2)]. Only aldolase class is without hapten.

et al., 1997b; Yang and Schultz, 1999). The somatic mutations all appear to contribute synergistically to binding, where the $\Delta\Delta G$ of binding for the forward mutation (somatic mutation introduced into the germline) is not equivalent to that of the reversion mutation (somatic mutation removed from the mature) (Yang and Schultz, 1999). This is the only system in which each somatic mutant has been studied in the germline background and the mature background.

An understanding of the 48G7 system has been facilitated by a thorough structural analysis. The results of this work are quintessential, in that the hapten-bound structures support a model of lock-and-key binding in mature antibodies and induced fit in germline antibodies. The mutations appear to constrain the germline conformation that binds ligand (Wedemayer *et al.,* 1997b).

Antibody 48G7 has undergone affinity maturation through somatic mutation, acquiring nine amino-acid substitutions relative to the original germline sequence. 48G7 binds a number of *p*-nitrophenyl phosphonates, including the transition state analog with which it was elicited, and catalyzes the hydrolysis of the corresponding nitrophenyl esters and carbonates with rate accelerations in excess of 10^4 over that of the uncatalyzed reaction (Table II) (Wedemayer *et al.,* 1997a). None of the nine mutations are in direct contact with the bound hapten; the closest is residue HisL55, which is located at a distance of 5.3 Å. Therefore, the improvement in binding affinity seen during the 48G7 affinity maturation process is due to generalized mid- or long-range effects propagated to the hapten binding site. The functional significance of such changes relative to the corresponding changes in structural details for the germline versus affinity-matured antibodies is discussed below.

The crystal structures of 48G7 have been determined in the presence of hapten at 1.95 Å resolution and in the absence of hapten at 2.7 Å resolution (Patten *et al.,* 1996; Wedemayer *et al.,* 1997a; Wedemayer *et al.,* 1997b). The root-mean-square difference between the two structures is 0.30 Å for the variable domains and 0.45 Å for the constant domains. Comparison of the two active sites shows that no significant changes occur on hapten binding to 48G7, as any main chain and side chain displacements are negligible. There are also no significant changes, in either the packing of V_H and V_L or the positions of the key active site residues. These key residues include ArgL96, TyrH33, and HisH35, which hydrogen-bond to the phosphonyl oxygens of the hapten; TrpH47, SerL93, TyrL94, and ArgH50, which fix the orientation of the three oxyanion binding residues; the TyrH98, TyrH99, TyrL91, and LeuL89, which form van der Waals contacts with the hapten.

Table II

Catalytic Antibody Affinity Maturation Kinetics

Antibody	Number mutations	k_{cat} (min^{-1})	K_m (μM)	$k_{\mathrm{cat}}/k_{\mathrm{uncat}}$	K_d (nM)
Germline 48G7	—	<1	not available	not available	135,000
48G7	9	5.5 (ester)	391 (ester)	1.6×10^4	4.5
Germline 39A11	—	10.2	1400 (diene)/450 (dienophile)	not available	379 (hapten)
39A11	2	40.2	1200 (diene)/740 (dieneophile)	not available	10 (hapten)
Germline 7G12	—	not available	not available	not available	not available
7G12	5	0.41	150 (Cu(II))	not available	not available
Germline AZ-28	—	0.8	73	163,000	670
AZ-28	6	0.023	74	5300	17

Legends for Color Insert

FIG. 2. Ribbon superpositions of the variable regions of the germline Fab-hapten complex (purple) and mature 48G7 Fab-hapten complex (red). Side chains for the somatic mutation sites are indicated in light green (germline) and dark green (mature). Mutations include SerL30 → Asn, SerL34 → Gly, AspL55 → His, GluH42 → Lys, GlyH55 → Val, AsnH56 → Asp, GlyH65 → Asp, AsnH76 → Lys, AlaH78 → Thr (from Wedemayer *et al.,* 1997b).

FIG. 3. Superpositions of the structures of the esterolytic germline Fab without hapten (blue) and the germline Fab-hapten complex (purple), illustrating the structural changes that occur on hapten binding to the germline antibody. All figures have the linker portion of the hapten omitted for clarity. Gray dashed lines denote hydrogen bonds in the hapten-free complex and black dashed lines denote hydrogen bonds in the hapten-bound complex (from Wedemayer *et al.,* 1997b). (a) The structural reorganization of CDR3 on hapten binding. The side chain of TyrH99 rotates to make room for the hapten, the side chain of TyrH98 inserts between TyrH99 and TyrH33, and TyrH33 moves toward the phosphonate group of the hapten. These movements establish new interactions, denoted with yellow dashed lines, that show the π-cation interaction between ArgL46 and TyrH99, the π-π interaction between TyrH99 and TyrH98, and the T-stack interaction between the aryl rings of TyrH98 and TyrH33. (b) The interactions between residues in CDRH1, CDRH2, and CDRL3 on hapten binding to the germline antibody. The guanidinium group of ArgH50 forms hydrogen bonds to the hydroxyl groups of TyrH33 and TyrL94 on hapten binding, and is additionally positioned by a hydrogen bond to AsnH56. (c) Close-up view of the phosphonate center in the combining site, illustrating the germline residues directly involved in hydrogen bonding with the hapten (side chains from HisH35, TyrH33, and ArgL96). TyrH33 moves 2.2 Å toward the phosphonate group, occupying approximately the same position as it does in the affinity-matured hapten-bound structure.

FIG. 4. Superpositions of the structures of the germline Fab-hapten complex (purple) with the mature 48G7 Fab-hapten complex (red), illustrating the changes that occur during affinity maturation. All figures have the linker portion of the hapten omitted for clarity. Gray dashed lines denote hydrogen bonds in the germline complex and black dashed lines denote hydrogen bonds in the affinity-matured structure (from Wedemayer *et al.,* 1997b). (a) Reorganization of CDR3 on hapten binding. The side chain of TyrH99 has rotated ~90°, forming a double T-stack arrangement between TyrH99, TyrH98, and TyrH33, denoted with yellow dashed lines. The additional hydrogen bond between the hydroxyl of TyrH33 and the phosphonate oxygen of the hapten is also shown. (b) The somatic mutations GlyH55 → Val and AsnH56 → Asp (shown in green) reorganize the CDRH1, CDRH2, and CDRL3 regions, with a series of interactions that influence residues in direct contact with the hapten molecule. Importantly, hydrogen bonds between ArgH50, TyrL94, and TyrH33 stabilize the orientation of the TyrH33 side chain. (c) In the affinity-matured Fab-hapten complex, an additional hydrogen bond is formed between the hydroxyl residue of TyrH33 and the phosphonate group of the hapten. Additionally, in the affinity-matured complex, the side chains of ArgL96 and HisH35 are closer to the phosphonate group of the hapten. (d) View of the combining site illustrating the potential steric interaction at position L34. Somatic mutation of position L34 (SerL34 → Gly) removes the side chain hydroxyl group from potential interference with hapten binding in the observed orientation for the affinity-matured complex. (e) The somatic mutations AsnH76 → Lys and AlaH78 → Thr (side chains shown in green) control the orientation of CDRH1, with a variety of hydrogen bonds and packing interactions for the germline and affinity-matured Fab-hapten complexes.

FIG. 5. Ribbon superpositions of the variable domains of the mature (red) 39A11 hapten-bound and the germline (blue) hapten-free structures. Positions of somatic mutation are indicated (ValL27c → Leu and SerL91 → Val) (from Romesberg *et al.,* 1998b).

FIG. 6. Close-up view of the superimposed mature 39A11 hapten-bound combining site (blue) and the hapten-free germline form of the antibody (red). No significant structural changes at the combining site are evident (from Romesberg *et al.,* 1998b).

FIG. 7. Ribbon diagram of the hapten-bound variable domain of 7G12. The heavy chain is shown in blue and the light chain is shown in red. Somatic mutations are colored in green (from Romesberg *et al.,* 1998a).

FIG. 8. Close-up of the combining site of 7G12. Van der Waals surfaces for key residues are shown. The prophyrin hapten is shown in yellow, the heavy chain is shown in blue, light chain is shown in red, and positions of somatic mutation are shown in green (from Romesberg *et al.,* 1998a).

FIG. 9. (a) Overlay of the combining site for AZ-28 affinity-matured structures, with (purple) and without (blue) bound hapten. Hapten is shown in yellow, with oxygen atoms colored red and nitrogen atoms colored blue (from Mundorff *et al.,* 2000). (b) Overlay of the combining site for the germline antibody structure, with (gray) and without (green) bound hapten. Hapten is shown in blue, with oxygen atoms colored red (from Mundorff *et al.,* 2000). (c) Close-up view of the interaction between positions TyrH100a and L34 (SerL34 → Asn) in the antibody combining site. Four different structures are included in the overlay. The AZ-28 affinity-matured Fab structures are colored purple (hapten bound) and blue (hapten free), and the germline Fab structures are colored gray (hapten bound) and green (hapten free) (from Mundorff *et al.,* 2000).

FIG. 10. Side and top down views of the superimposed structures of the affinity-matured Fab hapten-bound complex (purple) and the germline hapten-bound complex (grey). Hapten is shown in blue (from Mundorff *et al.,* 2000).

FIG. 11. Superimposed close-up view of the antibody combining sites, showing the altered orientations of the hapten in the germline and affinity-matured complexes. The AZ-28 affinity-matured Fab structures are colored purple (hapten bound) and blue (hapten free), and the germline Fab structures are colored grey (hapten bound) and green (hapten free).

FIG. 12. Ribbon diagram of wild type *p*NB esterase, looking down into the active site. Loops that are not visible in the electron density are shown with dashed lines, and loops that reorganize most significantly are shown in gold. The catalytic triad is shown in red, and secondary structures are labeled. (Figure from Spiller *et al.,* 1999.)

FIG. 13. Ribbon diagram of mutant 5-6C8, looking down into the active site. Loops that reorganize most significantly are shown in gold. The catalytic triad is shown in red, and mutations are shown in light blue. (Figure from Spiller *et al.,* 1999.)

FIG. 14. Ribbon diagram of mutant 8G8, looking down into the active site. Loops that reorganize most significantly are shown in gold. The catalytic triad is shown in red, and mutations are shown in light blue. (Figure from Spiller *et al.,* 1999.)

In contrast to the lock-and-key fit of hapten binding to 48G7, binding of hapten to the germline antibody leads to significant structural changes (Wedemayer *et al.,* 1997b). With respect to the liganded and unliganded germline antibody Fab structures, the root-mean-square deviation for all the C_α atoms of the variable region is 0.60 Å and the V_LV_H domain orientation changes by 4.6°. Whereas there are sixteen hydrogen bonding, electrostatic, and van der Waals interactions between the heavy and light chain domains in the free germline Fab, there are twenty-six such interactions in the germline Fab-hapten complex. These gross structural changes are accompanied by significant reorganization of the combining site residues (Figs. 2 and 3, see color insert). Binding of hapten leads to a repositioning of the active site residue TyrH33, along with the formation of a network of three hydrogen bonds between the side chains of TyrH33, ArgH50, and TyrL94. In the mature antibody, these interactions are reinforced by the AsnH56 → Asp and GlyH55 → Val somatic mutations, which fix the position of the TyrH33 hydroxyl group (the proposed oxyanion hole in 48G7) in the mature antibody combining site. Additional combining site repositioning observed on hapten binding to the germline antibody includes residues TyrH33 and TyrH98 moving 5 Å closer together into a T-stack arrangement, with an interaction between TyrH98 and the aliphatic linker of the hapten. Further repositioning involves an approximately 90° rotation of the TyrH99 side chain, resulting in the formation of a double T-stack arrangement between the side chains of TyrH99, TyrH98, and TyrH33, plus an additional π-cation interaction between the aryl ring of TyrH99 and the ArgL46 side chain; ArgL46 is oriented by an interaction with the side chain of AspL55 (an additional position of somatic mutation) (Fig. 3a). In total, as shown in Figures 3 and 4 (see color insert), the structural changes that occur in the germline-hapten complex become preorganized in the combining site of the mature antibody, consistent with the correspondingly increased binding affinity and improved kinetics for 48G7.

B. Diels Alderase (39A11) Antibody Maturation

One of the most important facets of catalytic antibodies is developing antibodies that catalyze reactions with no enzymatic counterparts. Pericyclic reactions occupy an important niche in organic synthesis methodology; examples include cycloaddition reactions, sigmatropic rearrangements, and electrocyclic ring closure reactions. The first such reaction to be successfully attained using a catalytic antibody system was for a Diels-Alder cycloaddition reaction. Catalytic antibody 39A11 catalyzes the cycloaddition of a diene and dieneophile to generate a Diels-Alder adduct (Romesberg *et al.,* 1998b). A bicyclo [2.2.2] octene hapten was used to mimic the boatlike $4\pi + 2\pi$ electron system of the transition state for this pericyclic reaction. Hapten was intentionally designed to destabilize bound product, which prefers a chair-like conformation (Braisted and Schultz, 1990). As expected, the boat-like transition state analog leads to antibodies that do not exhibit product inhibition.

The structure of the germline form of antibody 39A11, as well as the affinity-matured form of the antibody with bound hapten have been determined (Fig. 5, see color insert). (Romesberg *et al.,* 1998b). The overall structural features for the two antibody molecules are very similar, with C_α positions of V_H and V_L having minimal root-mean-square deviations (0.45 and 0.51 Å, respectively). Small differences are observed for the side chain of PheL87, removed from the active site, and for the position of residue L91 relative to HisL34 (SerL91 is mutated to ValL91 in the mature form of the antibody, with a distance of 4.2 Å between ValL91 and the hapten). It is interesting to note that neither somatic mutation nor ligand binding seems to result in any substantial structural or conformational changes in the active site structure.

Well-defined electron density was observed for the hapten molecule in the 39A11 Fab structure. The hapten is bound in a ''saddle'' fashion on the surface of the antibody combining site, with 79% of the hapten molecule buried within the Fab cleft (Fig. 6, see color insert). There are eighty-nine van der Waals interactions and three hydrogen bonds between the hapten and heavy chain of the antibody. The bicyclo[2.2.2]octene moiety of the hapten is buried in a hydrophobic pocket, making contacts with numerous heavy chain residues. The *N*-phenyl substituent of the hapten is packed between the backbone atoms of GlyH33, GluH96, and the methylenes of ArgH97. The carbonyl oxygen atom of the carbamate substituent at the bridgehead position of the hapten (corresponding to the C-1 substituent position of the analogous diene substrate) is coordinated through a water-mediated hydrogen bond to the Nε1 position of TrpH50. Additionally, TrpH50 is involved in a π-stacking interaction with the succinimido moiety of the hapten (which would correspond to the maleimide group of the substrate dienophile). A further interaction between the hapten and Fab is observed between the succinimido carbonyl group anti to the carbamate moiety in the hapten and the AsnH35 side chain; this interaction is analogous to Lewis acid catalysis in non-enzymatic Diels-Alder cycloadditions. AsnH35 is also oriented by a hydrogen bond between the carboxamide oxygen and the side chain of TrpH47.

Overall, the structure of the 39A11 hapten complex is consistent with the requirements for successful Diels-Alder reactivity, suggesting that the antibody would bind the diene and dieneophile in a similar fashion, leading to a reactive orientation for both substrates with reduced translational and rotational degrees of freedom. The diene would be bound in the hydrophobic pocket of 39A11 in close proximity to the dienophile, with the position of the carbamate substituent fixed by the water-mediated hydrogen bond to TrpH50. The dieneophile would be oriented by hydrogen bonding and π-stacking interactions with the maleimide ring. In addition, interaction of the maleimide group with AsnH35 would potentially make the olefin more electron-deficient, leading to a more reactive dienophile moiety in the combining site.

The crystal structures of the germline and affinity-matured forms of 39A11 lend a possible explanation to the observed increase in binding affinity and catalytic activity for the affinity-matured antibody. Mutation of the active site residue ValL91 (from SerL91) yields a binding pocket with increased hydrophobicity. Importantly, the mutation SerL91 → Val is at the base of the combining site, 4.2 Å from the hapten binding site. The portion of the hapten that is closest to site L91 is hydrophobic and the crystal structures show packing between hapten and protein that is too close to permit solvation at this site. Thus, the serine is desolvated in the liganded state of the germline antibody. Somatic mutation removes this unfavorable interaction by mutating position L91 to valine. However, the unliganded state is not sufficiently destabilized by the mutation, and the net effect is tighter binding in the mature versus the germline form of 39A11. These different constraints on the active site for 39A11 might lead to an increase in the rate as a result of improved packing interactions with the kinetically favored *endo* transition state.

C. *Metal Chelatase (7G12) Antibody Maturation*

Antibodies were generated against a bent *N*-alkylmesoporphyrin, which is a known inhibitor of the enzyme ferrochelatase. Ferrochelatase catalyzes the insertion of Fe(II) into protoporphyrin, as part of the heme biosynthetic pathway (Lavallee, 1988). Antibody 7G12 was found to catalyze the insertion of divalent metal ions into porphyrins, with a rate similar to that found for ferrochelatase (Cochran and Schultz, 1990). The structural data presented below supports the hypothesis that the transition state for porphyrin metallation involves a distortion of the macrocyclic ring system to facilitate metal insertion, and that this bent porphyrin acts as a good transition-state mimic in the development of catalytic antibody 7G12.

The three-dimensional crystal structure of the complex between the Fab fragment of 7G12 and *N*-methylmesoporphyrin was determined

(Romesberg *et al.,* 1998a), and the porphyrin hapten is bound in an ''inserted'' orientation in the antibody combining site (Fig. 7, see color insert). There are 114 pairwise interactions between the hapten and the antibody, predominantly hydrophobic contacts between the porphyrin and residues TyrL36, LeuL46, TyrL49, TyrL55, GlnL89, TyrL91, TyrL94, LeuL96, TrpH33, MetH50, and MetH97 (Fig. 8, see color insert).

Packing interactions of the hapten with residues TyrL49 and TyrL91 appear to play an important role in binding the distorted conformation of the *N*-methylmesoporphyrin. TyrL91 π-stacks with pyrrole ring B, which is coplanar with pyrrole rings C and D of the hapten molecule. TyrL49 π-stacks on pyrrole ring A, which is distorted out of the plane of the other three pyrrole rings by approximately 42°. Packing interactions between these two tyrosine side chains and the porphyrin pyrrole rings are reinforced by interactions with other residues. TyrL91 and TyrL49 contact the side chains of ProL32 and AsnL53, respectively, and TyrL49 makes an additional contact, forming a hydrogen bond between the aryl oxygen atom and the side chain of AsnL53. Pyrrole rings B, C, and D are packed tightly against the side chains from ArgH95, AspH96, and MetH97. Residues TyrL91 and MetH97 contact opposing faces of pyrrole ring B, possibly acting as a ''clamp'' to fix the position of this ring.

An additional important structural feature of the 7G12 hapten bound structure is the positioning of the carboxylate side chain from AspH96 over the central cavity of the porphyrin (1.9 Å from the center of the porphyrin ring). The orientation of this carboxyl group is further reinforced by a stabilizing hydrogen bond from ArgH95 to the Oδ2 position of AspH96, presumably ''locking'' the carboxylate residue in place. However, it is unknown whether AspH96 is acting as a general base to remove protons from the prophyrin ring during metal insertion, or whether the carboxylate is participating in the metal insertion reaction by forming a direct metal carboxylate bond. In the latter case, it is possible that the hydrogen bond between AspH96 and ArgH95 could act to reduce the entropic cost of losing rotational degrees of freedom upon metal ion coordination.

There are several structural similarities between antibody 7G12 and the ferrochelatase from *B. subtilis* (Al-Karadaghi *et al.,* 1997). Both proteins contain a hydrophobic binding cleft formed by hypervariable loops between two similar domains, with critical hydrophobic residues contacting the prophyrin rings (Trp230 for ferrochelatase, TyrL49 and TyrL91 in 7G12). Porphyrin binds edge-on in the cleft of 7G12 with its proprionic groups oriented into solvent and the A and B rings buried in the binding pocket; similar binding is predicted to occur with ferrochelatase. Resonance Raman experimental results are consistent with the ferrochelatase enzyme distorting the porphyrin ring in a doming fashion. A porphyrin ring distortion is also observed for 7G12 binding; however, Resonance Raman results are consistent with an alternating up-and-down tilting of two opposing pyrrole rings due to antibody binding (Blackwood *et al.,* 1998). Both ferrochelatase and 7G12 bind metal ions with an active site residue, Nε2 of His and Oδ1 of Asp, respectively. Finally, there is a possible difference between the two catalysts in that a basic residue may act as a general base in ferrochelatase, deprotonating the porphyrin prior to the metal insertion step. In 7G12 there is no obvious basic residue present in the active site for analogous catalysis; however, the carboxylate side chain of AspH96 possibly acts to deprotonate the porphyrin ring and might possibly be involved in transient metal coordination during the insertion process.

Somatic mutation to 7G12 resulted in five amino acid substitutions. It was not possible to differentiate between two alternate V_H germline genes, differing at a single amino acid in the coding region (Asn or Arg at residue 50), which were somatically mutated to MetH50 in 7G12. There are two additional mutations in the 7G12 V_H chain, SerH76 → Asn and SerH97 → Met. 7G12 also has two somatic mutations in V_L, SerL14 → Thr and AlaL32 → Pro. MetH97 and AsnH76 are 3.4 Å and 21 Å from the hapten, respectively, while residues ThrL14 and ProL32 are 29 Å and 7.6 Å from the hapten, respectively. The mechanistically important residues TyrL49, TyrL91, AsnL53 and ArgH95 are encoded in the germline DNA. AspH96 in CDRH3 is encoded by a random non-palindromic insertion of nucleotides; therefore, it was not possible to determine whether the origin of this residue was in the germline DNA or arose from a somatic mutation.

Analysis of the structure of the 7G12 Fab-*N*-methylmesoporphyrin complex suggests that residues ProL32 and MetH97, which were introduced during affinity muturation, play a central role in distorting the porphyrin substrate toward the reactive, bent conformation. To test this notion, the contributions of these mutations to catalysis were examined by generating the individual ProL32 → Ala and MetH97 → Ser mutants by site-directed mutagenesis (Romesberg *et al.,* 1998a). This work indicated that the ProL32 → Ala mutant possessed a virtually unchanged k_{cat} of 0.37 min^{-1} but a significantly increased K_M of 400 μM. These findings are consistent with the interpretation that the contribution of ProL32 to substrate binding is not differentially manifested in the ground and transition states. This result is also consistent with the suggstion that ProL32 plays a role in stabilizing the TyrL91 interaction with the B ring of the pyrrole. The MetH97 → Ser mutant was able to catalyze the insertion of Cu(II) into mesoporphyrin, but with a reduced k_{cat} of

0.022 min^{-1} and a K_M of 191 μM. Thus, a single somatic mutation at position H97 increased k_{cat} by more than an order of magnitude while leaving K_M virtually unchanged. The fact that the SerH97 and MetH97 mutant antibodies have similar binding affinities for the planar substrate, but differ in affinity for the distorted transition state indicates that the strain induced at the A and B pyrrole rings is manifested largely in the transition state. Overall, in a fashion similar to that observed for 48G7, the majority of hapten binding residues are encoded by germline amino acids, with additional stabilization afforded by somatic muations near or close to the antibody combining site.

D. Oxy-Cope (AZ-28) Antibody Maturation

The monoclonal antibody AZ-28 catalyzes the isomerization of (−)-chorismate to prephenate, in a similar fashion to the reactivity observed for the naturally occurring chorismate mutase enzyme. Chorismate mutase catalyzes a pericyclic reaction involving a 3,3-sigmatropic rearrangement, which is a key reaction step in plants and bacteria for biosynthesis of aromatic amino acids (Gray *et al.,* 1990). Catalytic antibody AZ-28 was elicited using a bicyclic hapten designed to be a structural analog of the asymmetric chair-like transition state geometry believed to be operative during the oxy-Cope rearrangement.

Contrary to all other catalytic antibody systems, affinity maturation results in tighter binding (670 to 17 nM) but decreased catalytic rate. The most catalytically active affinity matured oxy-Cope antibody is AZ-28 ($k_{\text{cat}}/k_{\text{uncat}}$ = 5300 versus germline $k_{\text{cat}}/k_{\text{uncat}}$ = 163,000), which undergoes six somatic mutations on maturation. Most of the loss in catalysis is due to the single somatic mutation, Ser L34 → Asn, which reduces $k_{\text{cat}}/k_{\text{uncat}}$ to 16,000 in the germline background (Ulrich *et al.,* 1997). This system has been studied by a combination of structural, mutagenesis, binding affinity, and kinetic methods (Ulrich *et al.,* 1997). A likely explanation for this decrease in catalytic activity on affinity maturation is provided by the crystal structure of AZ-28 in the presence of hapten. As described below, the structure reveals that the somatic mutation SerL34 → Asn restricts the rotation of the hapten phenyl rings, with respect to the cyclohexyl group, to a catalytically unfavorable perpendicular conformation (Ulrich *et al.,* 1997). A co-planar arrangement of the phenyl rings with the delocalized transition state lowers the free energy of activation by hyperconjugation (Fleming, 1978). Thus, maturation resulted in reduced catalysis.

The three-dimensional crystal structure was solved for the AZ-28 antibody Fab complexed with hapten (Ulrich *et al.,* 1997; Ulrich and Schultz, 1998). In this structure, the hapten is bound in an ''inserted'' orientation in a deep, cylindrical cavity about 8.3 Å wide and 18.5 Å deep (Fig. 9, see color insert, Mundorff *et al.,* 2000). The hapten is bound in a chair-like geometry with the hydroxyl and aryl substituents of the cyclohexyl ring located equatorially. Both phenyl substituents make extensive contacts with active site residues; the 5-phenyl group is buried at the bottom of the combining site cavity and makes contacts with TrpH47, TrpH103, PheL36, LeuL89, PheL98, AlaH93, and GluH35, and the 2-phenyl substituent is located near the surface of the binding pocket, making a π-stacking interaction with the imidazole ring of HisH96 and a van der Waals contact with TyrL91. The cyclohexyl ring of the hapten, which mimics the cyclic $4\pi + 2\sigma$ transition state for the pericyclic reaction, is rotated out of the planes of the 5- and 2-phenyl rings by 81° and 85°, respectively. The position of the cyclohexyl ring is fixed by hydrogen-bonding interactions between the hydroxyl group at C-1 of the hapten and the imidazole NH and backbone NH groups of HisH96, as well as water-mediated interactions with the carboxylate group of GluH50 and the side chain of TyrL96. An additional water-mediated hydrogen bond is formed between the hapten amide group and TyrL91. Finally, numerous van der Waals contacts are present between the cyclohexyl ring of the hapten and side chains from TyrL91, TyrL96, GlyH95, TyrH100a, AsnL34, and AspH101.

The X-ray crystal structure for AZ-28 has a variety of structural features that are consistent with the proposed mechanism operative for the oxy-Cope rearrangement. The antibody binds the transition stage analog in a chair-like conformation, consistent with the preferred chair transition state for this pericyclic reaction (Doering and Roth, 1962). The positions of the C-2 and C-5 atoms are fixed in the antibody-bound hapten molecule; in a similar fashion, the C-2 and C-5 positions in the hexadiene substrate should be held in a fixed position by conserved van der Waals interactions locking in the two phenyl substituents in the antibody combining site. This bound conformation of the acyclic ($4\pi + 2\sigma$) system of the hexadiene substrate should enforce a molecular conformation close to the transition state for the rearrangement reaction, consistent with the catalysis observed for AZ-28.

The crystal structures of AZ-28 and the related germline precursor (Fig. 9) (Mundorff *et al.,* 2000) provide an explanation for the inverse correlation between hapten binding affinity and catalytic rate. During affinity maturation, an increase in binding interactions between the hapten and antibody fix the orientation of the phenyl substituents with respect to the cyclohexyl ring. As a result of this tighter binding in mature AZ-28, the hexadiene core of the substrate rotates with respect to the phenyl substituents. This rotated orientation serves to increase hydrogen-bonding and packing interactions, making it less likely that maximal orbital overlap is achieved in the mature antibody, in contrast to the germline antibody, where hapten is bound less tightly. Indeed, the 5- and 2-phenyl rings are found to be rotated only 63.2° and 57.9°, respectively, in the germline · hapten complex. In the affinity matured AZ-28 · hapten complex, AsnL34 is located 3.7 Å from the cyclohexyl ring and, along with TyrH100a and AspH101, helps fix the conformation of the bound hapten (Figure 9c). Introduction of the somatic mutation SerL34 → Asn in AZ-28 leads to a modest increase in binding affinity (relative to the germline precursor G-28), but also affects the relative orientation of the diene and phenyl orbitals. This change in orbital alignment between the phenyl rings and the $4\pi + 2\sigma$ system of the diene substrate is consistent with the decreased rate observed for AZ-28.

The structural differences between the germline and mature antibody structures are small (Fig. 10, see color insert). The three most significant differences are in CDRH3, at the site of the catalytically important somatic mutation SerL34 → Asn, and in the conformation of the hapten itself (Fig. 11, see color insert). Residues H97 to H100a are in a similar conformation in both the unliganded and liganded affinity-matured structures. In contrast, this region is shifted 4.9 Å out of the active site in the germline antibody · hapten complex versus the unliganded germline structure. The outward displacement of CDRH3 in the germline · hapten complex relative to the affinity matured antibody · hapten complex also probably contributes to the decreased dihedral angles in the germline structure for the hapten 2- and 5-phenyl substituents.

The shorter hydrogen bond in the affinity-matured antibody · hapten complex (the L34 to Tyr-H100a distance is 2.6 Å in the affinity-matured antibody and 3.5 Å in the germline · hapten complex) appears to stabilize the conformation of CDRH3, but presumably fixes the substrate in a catalytically unfavorable conformation in the affinity matured antibody. In contrast, increased flexibility of CDRH3 in the germline structure would lead to a more open active site structure in the germline antibody · hapten complex. This would lower the rotational barrier around the 2-phenyl-C2 bond and allow for increased π-orbital overlap in the transition state. The net result would be an increase in catalytic rate for the germline antibody relative to the affinity-matured antibody. At the same time, this conformational flexibility is most likely to be responsible for the weaker binding affinity for hapten in the germline antibody.

Conformational analysis of the AZ-28 bound hapten suggests that it is not a true mimic of the oxy-Cope transition state. The lowest energy conformation for the hapten is with the cyclohexyl moiety in the chair-like configuration with the phenyl and hydroxyl moieties equatorial. In the lowest energy conformation, the phenyl rings are both roughly perpendicular to the cyclohexyl ring. A true mimic of the oxy-Cope transition state would have the phenyl rings coplanar to the cyclohexyl ring in order to maximize overlap of the phenyl π-orbitals with the $4\pi + 2\sigma$ orbitals of the Cope transition state. This orientation should decrease the activation free energy for the reaction and result in a higher rate of catalysis. Affinity maturation of the Cope system results in an increased binding affinity for the lowest energy conformation of the hapten, which in this case also happens to be the conformation least favorable for catalysis. The germline precursor has a lower binding constant because of the more open active site; this is characteristic of pre-affinity matured antibodies. However, unlike other catalytic antibody systems, this open active site results in a significantly higher rate of catalysis, which stems from the greater torsional freedom of the phenyl rings and greater probability of catalytically favorable π-orbital overlap.

Structural studies of the oxy-Cope catalytic antibody system reinforce the idea that conformational dynamics of both protein and substrate are intimately intertwined with enzyme catalysis, and consideration of these dynamics is essential for complete understanding of biologically catalyzed reactions. Indeed, recent single molecule kinetic studies of enzyme-catalyzed reactions also suggest that different conformations of proteins are associated with different catalytic rates (Xie and Lu, 1999). In addition, a number of enzymes are known to undergo conformational changes on binding of substrate (Koshland, 1987) that lead to enhanced catalysis; two examples are hexokinase (Anderson and Steitz, 1975; Dela-Fuente and Sols, 1970) and triosephosphate isomerase (Knowles, 1991).

Due to the nature of crystallographic studies, dynamic conformational changes that play an important role in accommodating changes in protein-substrate interactions along complex reaction pathways can only be alluded to indirectly through analysis of static structures. Dynamic studies on the oxy-Cope system could illuminate the model of catalysis suggested by this X-ray crystallographic analysis.

III. Structural Studies of Enzyme Directed Evolution

The previous sections described structural studies of antibody maturation as a method to understand the evolution of binding (and catalysis) in the immune system. The technique of directed evolution parallels the process of affinity maturation. Both methods use random mutagenesis and gene shuffling, followed by screening and/or selection to identify mutants with the desired function. In contrast to affinity maturation, directed evolution is not confined to one protein family—any number of gene fragments with regions of homologous DNA can be mutated and recombined. In both methods, however, similarly impressive results are obtained: a few amino acid mutations give rise to evolved proteins with distinctly improved properties. The remainder of this chapter presents a structural analysis of directed evolution products and offers insight into the causes of thermostability and altered substrate specificity.

A. *Structural Analysis of Adaptive Mechanisms to Evolutionary Pressures*

Directed evolution is a powerful technique that can be used to simultaneously identify areas in sequence, structure, and functional space that are essential in defining a protein's stability and function. Using low mutation rates and recombination to minimize the number of silent or deleterious mutations (Stemmer, 1994; Arnold, 1998), followed by selection or screening against environmental or biochemical challenges, directed evolution methods can quickly produce mutant proteins that are adapted for activity in a nonnative environment. These mutants can be instructive models in studies aimed at understanding protein evolutionary mechanisms and the adaptations necessary to maintain activity in nonnative environments. Unlike comparisons of related proteins from different species, where the majority of the mutations arise from genetic drift (Kimura, 1982), sequence or structural alignments of proteins obtained by directed evolution facilitate the immediate identification of accessible sequence changes that also result in desired structural and functional changes. In addition, directed evolution can be used to investigate the mechanisms by which proteins adapt to different selection pressures. The following discussion of *p*NB esterase divergent evolution represents a rare example in which the structural aspects of these adaptive mechanisms can be elucidated.

The enzyme *p*-nitrobenzyl (*p*NB) esterase has been evolved in two different experiments to obtain mutant forms of *p*NB esterase that maintain activity in two very different unnatural nonnative environments (Moore and Arnold, 1996; Moore *et al.,* 1997; Giver *et al.,* 1998; Gershenson *et al.,* 2000). Five generations of directed evolution with screening for increased activity in organic solvents led to the isolation of mutant 5-6C8, which contains seven amino-acid substitutions (Moore and Arnold, 1996; Moore *et al.,* 1997). The evolved enzyme has a hundred-fold increased esterase activity relative to wild-type (WT) enzyme in 25% dimethylformamide using the antibiotic *p*-nitrobenzyl loracarbef (*p*NB-LCN) as substrate (Moore *et al.,* 1997). Subjecting a first-generation organophile mutant to eight further generations of directed evolution, using screening for improved thermal stability, yielded mutant 8G8, which contains a total of thirteen amino-acid substitutions relative to wild type (Giver *et al.,* 1998; Gershenson *et al.,* 2000; see chapter by Wintrode and Arnold in this volume). Thermophile 8G8 has a 17°C increase in melting temperature (T_m = 69.5 °C), a 15°C increase in temperature for optimal activity (T_{opt} = 60°C), and increased activity toward the substrate *p*-nitrophenyl acetate (*p*NPA) relative to wild-type *p*NB esterase at 30°C. The crystal structures of the wild type, organophile 5-6C8, and thermophile 8G8 esterases have been solved (Spiller *et al.* 1999), allowing for the elucidation of the structural adaptive mechanisms invoked to stabilize *p*NB esterase under different environmental challenges. In total, these structures demonstrate how complex networks of interacting mutations can increase the stability and alter the functionality of *p*NB esterase.

B. Wild-Type pNB Esterase Structure

*Bacillus subtilis p*NB esterase is a member of the α/β hydrolase fold family (Moore and Arnold, 1996; Ollis *et al.,* 1992). The "canonical" α/β hydrolase fold consists of a mostly parallel eight-stranded β sheet surrounded on both sides by α helices (Nardini and Dijkstra, 1999). *p*NB esterase contains 489 amino acids arranged in a central thirteen-stranded β sheet that is surrounded by fifteen α helices (Fig. 12, see color insert). Similar to the structure of acetylcholine esterase (Sussman *et al.,* 1991), a large fraction of the *p*NB esterase structure has no defined secondary structure (52% random coil, 33% α helix, and 14% β sheet). This high degree of random coil structure is allowed in the α/β hydrolase fold, where large insertions in loops were found to be tolerated while still maintaining catalytic activity (Nardini and Dijkstra, 1999).

Sequence alignment of *p*NB esterase indicates a catalytic triad consisting of Ser189, His399, and Glu310, with the active site located in a cavity with approximate dimensions 20 Å by 13 Å by 18 Å (Fig. 12). The entrance to the active site is formed by four loops that reorganize substantially during divergent directed evolution. Residues 66–74 and 414–420 form one side of the entrance to the *p*NB esterase active site, with residues 316–320 and 260–268 forming the other side. In the wild-type structure, loops 66–74 and 414–420 are not observed in the electron density. However, mass spectrometric analysis identified an intact protein, with only the N-terminal methionine deleted. Presumably, these two loops occupy multiple conformations in the wild-type *p*NB esterase. The active site cavity is comprised of residues 105–108, 189, 193, 215–216, 268–275, 310–314, 362–363, 371, and 399–400.

C. Structural Characterization of 5-6C8, an Evolved Organophilic pNB Esterase

As seen in Figure 12 and 13 (see color insert) there are no substantial changes between the wild type and evolved 5-6C8 esterase structures, consistent with a root-mean-square deviation of only 0.67 Å for 467 C_α positions. The β sheet region, along with the majority of the α helical regions remain fixed in space, acting as a structural framework for the portions of the secondary structure that do move as a result of the mutations introduced into the *p*NB esterase sequence. Major rearrangements of the C_α positions for the four surface loops that surround the active site cavity are observed in the evolved 5-6C8 structure. In addition, repositioning of α helices surrounding the active site cleft, along with rearrangement of the relative positioning of a series of hydrophobic side chains that form the sides of the active site cavity serve to reshape the geometry of the active site cleft present in 5-6C8. However, even with all the observed structural changes in 5-6C8, the positions and orientations of the catalytic triad are superimposable with that found in the wild-type structure.

The seven amino-acid substitutions present in 5-6C8 (Table III and Fig. 13) are located in various regions throughout the structure and not in proximity to the esterase active site, with only one mutation (P317S) located near the entrance to the active site cavity. However, the amino acid substitutions are located in regions of the three-dimensional structure that were found to decrease the flexibility of surface loops, allowing the enzyme to recover activity in aqueous organic solvents. At the same time, this set of mutations reorganizes the esterase active site cavity for improved substrate binding and catalysis of the *p*-nitrobenzyl and *p*-nitrophenyl ester substrates. The stepwise addition of mutations to the wild-type esterase, as observed during the five generations of directed evolution that yielded the organophile 5-6C8, caused reinforced stabilization of the four surface loops that form the entrance to the active site cavity, as well as reorganization of the active site cavity geometry.

The His322 → Arg mutation that arose during the first generation of directed evolution plays a major role in reorganization of two key active site loops, 265–275 and 315–324. The side chain of Arg322 rotates ~95° away from the position of the His322 side chain found in the wild type structure, pointing directly toward the 265–275 loop. Loop 265–275 deviates substantially from the position in the wild-type structure, shifting closer to helices 9 and 11 with formation of an Arg322/Asp268 hydrogen bond. Loop 315–324 and helix 9 also change position slightly, due to the cross-loop interaction of residue 322 with residues 268 and 267 of

TABLE III

Proposed Role of Mutations Introduced into 5-6C8 and 8G8

56C8 Mutations	Location in structure	Generation	Proposed role for 5-6C8 mutations
L144M	buried, surface loop	3	Reorients helix 6, which stabilizes the 265–275 loop. Interacts with loops also affected by M358V, H322R, and L334S.
H322R	surface, surface loop	1	Stabilizes 315–324 loop, forms hydrogen bond with Asp268. Moves 265–275 loop. Interacts with loops also affected by M358V, L144M, and L334S.
L334S	surface, helix 9	4	Stabilizes movement of helix 9 and 265–275 loop, but destabilizes helix 10. Interacts with loops also affected by H322R, M358V, and L144M.
M358V	buried, helix 11	2	Reorganizes 265–275 and 315–324 loops and helix 9; Phe314 and Phe271 shift to space previously occupied by longer Met358 side chain, shifting Leu362, the 258–275 loop, and helix 9. Interacts with loops also affected by L144M, H322R, and L334S.
I60V	buried, loop	4	Stabilizes 66–74 and 414–420 loops, allows 10 intramolecular hydrogen bonds including Tyr72/Ser114, Tyr72/Arg415, and Leu70/Arg415.
P317S	surface, surface loop	5	Perhaps stabilizes new orientation of 315–324 loop.
Y370F	buried, helix 11	1	Perhaps results in more favorable hydrophobic packing.

8G8 Mutations	Location in structure	Generation	Proposed role for 8G8 mutations
L144M	buried, surface loop	1	Reorients helix 6, which stabilizes the 265–275 loop. Interacts with loops also affected by M358V, H322Y, and A343V.
H322R	surface, surface loop	1	Repositions the 315–324 loop. Determines the interaction between 265–275 and 315–324 active site loops. Forms hydrogen bond with Ile270, allowing Ser323/Thr326 to hydrogen bond and stabilize helix 9. Interacts with loops also affected by M358V, L144M, and A343V.
R322C		3	
C322Y		4	

A343V	buried, helix 10	2	Improves hydrophobic packing between helices 9 and 10. Interacts with loops also affected by L144M, M358V, and H322Y.
M358V	buried, helix 11	1	Shifts 265–275 (smaller shift than 56C8) and 315–324 loops; Phe315 shifts to space previously occupied by longer Met358 side chain, shifting Leu362 and Ile270. Interacts with loops also affected by H322Y, L144M, and A343V.
I60V	buried, loop	1	Stabilizes 66–74 and 414–420 loops, allows 10 intramolecular hydrogen bonds including Tyr72/Ser114 and Ser71/Arg415.
T73K	surface, surface loop	7	Forms salt bridge with Glu74; also in 66–74 loop (which is unstructured in wild type esterase).
G412E	surface, helix 13	5	Forms salt bridge with Arg415 in 414–420 loop (which is unstructured in wild-type esterase), reinforcing I60V mutation.
L313F	surface, helix 8, active site cleft	5	T stacks with Phe314.
Y370F	buried, helix 11	1	Possibly results in more favorable hydrophobic packing.
A400T	surface, loop, active site cleft	8	Thr400 is in the active site, possibly stabilizes His399 (part of the catalytic triad). Packs against Met416, with shifted position due to G412E mutation; larger Thr400 fills space formed by 416 movement.
I437T	buried, helix 14	3	Forms intramolecular hydrogen bonds with Gln433 and water-mediated with Lys441, stabilizing helix 14.
A56V	surface, loop	8	No observable changes.
T459S	surface, loop	7	Both Thr459 and Ser459 hydrogen bond to His456; no observable changes.

loop 265–275. Additionally, the Arg322 side chain is involved with the formation of a binding pocket for a sulfate ion. The side chains of residues Lys267, Arg322, Ser323, and Thr326 all bind to the anionic sulfate center, further stabilizing the 265–275 and 315–324 loops.

A second mutation found during the first generation of directed evolution, Tyr370 → Phe, might not be stabilizing or catalytically enhancing by itself. Previous studies determined that the double mutant H322R/Y370F was expressed at higher levels but without increased specific activity (in aqueous media) relative to wild-type enzyme (Moore and Arnold, 1996). In addition, no appreciable structural changes occur in the vicinity of position 370 because of the replacement of a tyrosine side chain with a correspondingly larger phenylalanine side chain. However, enthalpic changes occur with mutation of residue 370. A hydrogen-bonding interaction between the side chain of Tyr370 and the peptide carbonyl of position 199 is lost with the Phe370 substitution, but hydrophobic interactions seem to be more important energetically at this position. In both the wild type and 5-6C8 structures, the side chain of residue 322 is located in a hydrophobic interior region of the protein, making the more hydrophobic Phe370 residue enthalpically favorable versus Tyr370 for this position. Site-directed mutagenesis confirmed the stabilizing effects of Tyr 370 → Phe when introduced as a single mutation into wild type *p*NB esterase (Giver *et al.,* 1998).

The second generation mutation Met 358 → Val plays an additional role in the reorganization and stabilization of the new conformations of the 315–324 and 265–275 loops in the 5-6C8 structure. As shown in Figure 13, position 358 is located in helix 11, which packs against loop 315–324. Replacement of the extended methionine side chain with the more compact valine side chain creates an internal cavity in the protein core. This space is filled by a repositioning of the Phe314 and Phe271 side chains toward the Val358 side chain, along with the Leu362 side chain rotating toward the active site cavity. Movement of Phe314 causes the repositioning of loop 315–324 and the following helix 9, as well as a restructuring of the active site cavity around residues 314, 362, and 271. The 2 Å movement (C_α of wild type relative to C_α of organophile) of Phe271 causes a further repositioning of a seventeen amino acid stretch (from residues 258–275), which forms a portion of the entrance (loop 265–275) and wall to the active site cavity.

The third generation mutation Leu144 → Met reinforces the new orientation of the 265–275 loop. Replacement of the smaller leucine side chain with the larger methionine side chain causes a repositioning of the Phe233 side chain, which rotates 50° around $\chi 1$ and 30° around $\chi 2$. Movement of Phe233 allows Leu262 to shift into the vacated space, resulting in a reorientation of helix 6 and further stabilizing the new orientation of the 265–275 loop.

The fourth generation Leu334 → Ser mutation reinforces the new orientations for the 265–275 and 315–324 loops and helix 9. Substitution of serine for leucine at position 334 accommodates the movement of Phe271. However, this reorganization produces steric strain between Phe271 and Ser334 that is relieved by movement of helix 9, subsequent repositioning of helix 10, and repositioning of Ala343, which causes the alanine side chain to be more solvent exposed. To minimize this effect, Ala343 adopts unfavorable phi-psi torsion angles.

The second mutation found in the fourth generation, Ile60 → Val, stabilizes the two surface loops (66–74 and 414–420) that are unstructured in the wild-type enzyme. Replacement of isoleucine at position 60 with the smaller valine side chain allows a major reorganization of the structure of loop 66–74, including the formation of a new hydrogen bond between the side chain of Tyr72 and the side chain of Ser114. Additionally, loop 414–420 becomes structured due to the Ile60 → Val mutation, with cross-loop hydrogen bonds forming between the side chain of Arg415 and the backbone carbonyls to Tyr72 and Leu70. In total, the alternate loop structures of residues 66–74 and 414–420 found in 5-6C8 provide ten additional hydrogen bonds. This new series of interactions allows the 66–74 loop to pack against residues 60–63 and 110–116. In the wild-type esterase structure, steric clashes between the side chains of residues Ile60 and Leu68 block the 66–74 loop from approaching the 60–63 and 110–116 loop regions, preventing this compact 5-6C8 loop conformation from forming.

The role of the fifth generation Pro317 → Ser mutation is unclear but may stabilize the new orientation of the 315–324 loop.

D. Structural Characterization of 8G8, an Evolved Thermophilic pNB Esterase

The most thermophilic variant of *p*NB esterase, 8G8, has only thirteen mutations compared to the wild-type esterase, making it 97% identical to the wild-type esterase sequence, with a root-mean-square deviation of only 0.44 Å between the two C_α backbone structures. As with the 5-6C8 organophile structure, the catalytic triads of 8G8 and wild-type *p*NB esterase are superimposable. This high sequence and structural identity, in conjunction with the availability of crystal structures for both the wild type and thermophile, affords an interesting opportunity to study the structural basis for thermostability. Thermophile 8G8 is the product of eight generations of directed evolution, screening for retention of activity towards *p*NPA after incubation at high temperature (Giver *et al.,* 1998; Gershenson *et al.,* 2000). The starting point for further directed evolution was a thermophilic esterase from a preexisting library (generated during the organophile study) that already contained the mutations Ile60 → Val, Leu144 → Met, His322 → Arg, Tyr370 → Phe, and Met358 → Val.

As with organophile 5-6C8, the mutations found in thermophile 8G8 serve to both reorganize the active site and stabilize surface loops. The thirteen mutations in 8G8 fall into four groups. Mutations in the first group (Leu144 → Met, Met358 → Val, His322 → Tyr, and Ala343 → Val) all interact to reorganize the 265–275 and 315–324 loops. A second group of mutations (Ile60 → Val, Gly412 → Glu, and Thr73 → Lys) stabilize the 66–74 and 414–420 surface loops. These loops, which do not contribute to the lattice structure in any of these isomorphous crystals, are unstructured in the wild type esterase. The third group of mutations (Tyr370 → Phe, Leu313 → Phe, Ile437 → Thr, and Ala400 → Thr) have individual, interpretable effects, while the fourth group of mutations (Thr459 → Ser and Ala56 → Val) do not have definitive explanations for their contributions to thermostability. The thirteen sites of mutation and their putative mechanisms of stabilization are summarized in Table III.

The first generation mutations found in thermophile 8G8 are positions of mutation that were shared with the organophile 5-6C8 (144, 358, 322, 60, 370). Mutations Leu144 → Met and Met358 → Val make similar contacts in both the 8G8 and 5-6C8 structures; however, the overall effect in the thermophile structure is greatly influenced by the mutation at position 322, as will be discussed. The mutation Met358 → Val allows the residues Ile270, Leu362, and Phe315 to move, but the large reorganization of helix 6 and loop 265–275 that was observed in the organophile structure is not found in the 8G8 structure. Movement of the 315–324 loop in the organophile structure is opposite that of the thermophile 8G8 structure relative to the wild type structure (compare Figs. 12, 13, and 14, see color insert). Deviation in the 315–324 loop between the organophile and thermophile structures is in the range of 2.9 Å (position of residue 322 in 5-6C8 versus 8G8). The explanation as to why identical mutations in 5-6C8 and 8G8 have different structural (and functional) results involves the pivotal role of residue 322, which differs in all three esterases. Position 322 dictates the nature of the interaction between the 315–324 loop and the 265–275 loop. In the evolution of thermophile 8G8, residue 322 starts out as an arginine in generation one, which then mutates to a cysteine in generation three, and finally in generation four mutates again to a tyrosine (Giver *et al.,* 1998). In the thermophile 8G8 structure, the 265–275 loop follows the wild-type loop conformation much more closely than observed in the organophile 5-6C8 structure, but the 315–324 loop is pulled closer to the active site in 8G8. Residue 322 moves 2.5 Å (measured relative to the C_α) from its wild-type position, and the C_β position points ~90° away from its wild-type position and towards the 265–275 loop. Unlike Arg322 (found in 5-6C8), Tyr322 (found in 8G8) does not interact directly with the adjacent 265–275 loop (except for a 3.5 Å hydrogen bond to the backbone amide of residue 270). The new position of the 315–324 loop observed in 8G8 allows Ser323 to hydrogen bond to Thr326, stabilizing both the loop and helix 9. The first-generation mutation Ile60 → Val is found to once again stabilize the 66–74 and 414–420 loops in 8G8, allowing the formation of a slightly different hydrogen bonding network in this instance: side chains from Tyr72/Ser114, and Ser71/Arg415 form hydrogen bonds, stabilizing both the 66–74 and 414–420 loops. The final first-generation mutation found in thermophile 8G8, Tyr370 → Phe, adopts the same positioning and effects as observed in the 5-6C8 organophile structure.

The second generation mutation found in 8G8, Ala343 → Val, is in helix 10, and most likely improves hydrophobic packing between helices 9 and 10. The third generation mutation Ile437 → Thr is in helix 14 and results in an intrahelix hydrogen bond to the main chain oxygen of position 433, stabilizing this helix. The other third generation mutation, Arg322 → Cys, mutates again during the fourth generation of directed evolution from Cys322 to Tyr322. This mutation was included in the previous discussion, due to the pivotal involvement of position 322 in the reorganization and stabilization of loops 265–275 and 315–324.

The fifth generation mutation Leu313 → Phe results in a stabilizing face-to-edge interaction with Phe314 (Hunter *et al.,* 1991; Serrano *et al.,* 1991). The orientation of Phe314 in the thermophile structure is rotated and shifted toward Phe313, relative to the orientation of Phe314 in the organophile structure. Hence, the shape of the wall in the active site cavity around positions 313 and 314 in 5-6C8 and 8G8 is slightly different, consistent with the altered substrate specificity observed in 5-6C8 versus 8G8. The second fifth generation mutation, Gly412 → Glu, further stabilizes loop 414–420, forming a salt bridge with Arg415, which stabilizes helix 13. This interaction has been shown by mutational analysis to contribute to the thermostability of 8G8 (Giver *et al.,* 1998).

The seventh generation mutation Thr73 → Lys is also in the loop stabilized by the substitution Ile60 → Val; due to disorder of loop 66–74, the position of Thr73 in the wild-type structure is unknown. In the thermophile 8G8, the substitution Lys73 forms a salt bridge with Glu74. Both of these stabilizing mutations (the fifth generation Gly412 → Glu mutation and the seventh generation Thr73 → Lys mutation) occur after the 66–74 and 414–420 loops are fixed (the loops are fixed in the first generation) and are thus clear examples of stabilizing salt bridges that could not be designed into the wild-type esterase structure, illustrating the power of iterative directed evolution methods in protein engineering.

Although residue 400 is visible in all three *p*NB esterase structures, the eighth generation mutation Thr400 in 8G8 packs against Met416, which is not visible in the wild-type structure. Met416 is shifted away from residue 400 in the thermophile structure relative to the organophile structure, due to the salt bridge introduced by the fifth generation Gly412 → Glu mutation. The movement of Met416 is accommodated by the larger threonine side chain. Furthermore, the Thr400 side chain is positioned in the active site and may help stabilize His399 (part of the catalytic triad).

As seen in the antibody maturation structural studies, not all mutations can be rationalized, even in cases when structures are available for comparison (Romesberg *et al.,* 1998b; Wedemayer *et al.,* 1997b). Thus, screening of mutant libraries can often identify interesting proteins, but neutral mutations do arise in these populations. The ability of recombination to remove neutral mutations (Stemmer, 1994; Arnold, 1998) (performed most recently after the seventh generation in the evolution of 8G8) presents a new tool to be enlisted in the study of somatic mutation in the immune system, as well as in molecular evolution studies. The seventh generation Thr459 → Ser mutation and the eighth generation Ala56 → Val mutation are two ambiguous mutations present in the thermophile 8G8; both are exposed to solvent and are more than 25 Å from the active site. In addition, the mutations at positions 459 and 56 appeared in the same generation as known stabilizing mutations and therefore may not be stabilizing by themselves.

IV. Conclusions

Directed evolution and antibody affinity maturation offer efficient routes to redesigning proteins for new functional characteristics. Adaptive mutations and well-defined selection pressures allow structural analysis of the evolved products to provide insights into the molecular basis of protein structure and function. It is interesting to note that the majority of mutations that were obtained in the present maturation and directed evolution experiments were located at positions away from the enzymatic active sites. Perhaps this is due to the inherent difficulty in retaining catalytic activity with most active site amino acid substitutions.

Most rational design approaches to protein engineering currently incorporate active site amino acid substitutions and successful results are much more difficult to obtain using this limited range of mutational space. Even conservative amino acid substitutions in the active site can have pronounced effects on the active site shape and chemistry. In light of this apparent difficulty with replacing active site amino acid residues, the results described above with antibody affinity maturation and enzyme directed evolution might not be too surprising.

In cases when both germline and mature forms of a catalytic antibody have been studied, a variety of lessons have been learned. For 48G7, somatic mutation led to an increase in binding affinity as well as an increase in catalysis. The germline antibody was found to change shape on hapten binding (induced fit), whereas the mature antibody did not (lock-and-key fit). For the Diels Alderase antibody 39A11, polyspecificity was observed for the germline precursor, and the maturation process increased affinity for the target while decreasing affinity for background molecules. The ferrochelatase mimic, 7G12, presented an example of strain in catalysis introduced during the maturation process, where the transition state mimic was found to bind in a distorted, nonplanar geometry that was also corroborated by Resonance Raman measurements. For AZ-28, an oxy-Cope catalytic antibody, affinity maturation led to an increase in binding affinity, but a decrease in catalytic rate. However, this reduction in kinetic efficiency was due to the use of a flexible hapten capable of altering conformation during the maturation process to a more kinetically disfavored geometry. The germline antibody active site was found to be more open while the mature form was fixed in a disfavored orientation.

Similarly, directed evolution is a powerful tool for developing novel enzymes. In the directed evolution of *p*NB esterase, protein stabilization is accomplished by improved core packing, helix stabilization, the introduction of surface salt bridges, and the reduction of flexibility in surface loops, all of which are proposed mechanisms by which naturally occurring proteins from thermophiles are stabilized (Facchiano *et al.,* 1998; Jaenicke and Bohm, 1998; Vogt *et al.,* 1997; Zavodszky *et al.,* 1998). Interestingly, in the case of *p*NB esterase, the stabilizing forces introduced by directed evolution generated point mutations that allowed the evolution of two mutant enzymes with new optimized functions in two divergently unnatural environments. Organophile 5-6C8 was found to efficiently hydrolyze the bulky ester substrate *p*NB-LCN in 25% dimethyformamide solution, whereas thermophile 8G8 was found to efficiently hydrolyze *p*NPA, a much smaller substrate, with an improved thermal stability and 15°C increased optimal temperature of reaction relative to wild-type *p*NB esterase.

The improvements in *p*NB esterase function required mutiple generations to evolve, with subsequent generations adding new stabilizing factors, as well as reinforcing stabilizing factors present from previous generations of directed evolution. The complexity of the changes brought about by the directed evolution of *p*NB esterase underscores the difficulty of the rational design problem. Directed evolution, in which mutations in the later generations interact with the results of previous mutations, can reveal non-additive mutation pathways. Hence, both antibody affinity maturation and directed evolution processes seem to employ non-additive mutational solutions. The introduction of interacting mutations in successive generations indicates that rational design will benefit from efforts to predict the effects of single mutations and then introduce later generation mutations in an iterative manner.

Recently, the structure of an aminotransferase directed evolution product was reported by Oue *et al.* (1999). These authors also found clusters of interacting mutations distant from the active site. The importance of substitutions distant from the active site has been observed previously by Arkin and Wells (1998), and also reinforces the inability of current protein engineering efforts to *a priori* predict locations of beneficial mutations. The ability of directed evolution to identify structurally and functionally important amino acid clusters in proteins is similar to the ability of the immune system to control the shapes of the complementarity determining regions of antibodies with a small number of critical sites (Chothia and Lesk, 1987). Both directed evolution and antibody affinity maturation point to the existence of mutational hotspots that influence function across long distances. For example, the starting point enzyme for the thermophilic esterase study was a prior generation organophilic esterase that contained five of the seven mutations found in 5-6C8. Their continued presence during evolution in the divergent extreme conditions of the thermophile study highlights the importance of these positions of mutation on overall improved enzyme function. This once again confirms the existence of evolutionary ‘‘hot spots’’ of mutation. Another supporting observation was illustrated in the case of catalytic antibody 7G12, where somatic mutations were found to directly pack against the porphyrin binding residues. However, a molecular level structural analysis will be necessary to delineate the specific structural changes that occurred on affinity maturation to 7G12.

The structural studies described in this chapter highlight the importance of mutations away from the active site. Importantly, both close and distant productive mutations in both affinity-matured and directed evolution generated enzymes were found to ''tweak'' the desired enzyme function, by interacting with key areas of the protein scaffolding. X-ray structural characterization of both the ''before'' and ''after'' evolved protein structures provided the first opportunity to point out the location of critical mutations and key areas of the protein scaffold that have been found to modulate and attenuate overall enzyme function. Future studies involving this same combination of techniques should shed further light on the difficult problems associated with *de novo* protein design.

References

Al-Karadaghi, S., Hansson, M., Nikonov, S., Jonsson, B., and Hederstedt, L. (1997). *Structure* **5,** 1501–1510.

Anderson, W. F., and Steitz, T. A. (1975). *J. Mol. Biol.* **92,** 279–.

Arkin, M. R., and Wells, J. A. (1998). *J. Mol. Biol.* **284,** 1083–1094.

Arnold, F. H. (1998). *Acc. Chem. Res.* **31,** 125–131.

Barbas, I., C. F., Heine, A., Zhong, G., Hoffman, T., Gramatikova, S., Björnestedt, R., List, B., Anderson, J., Sture, E. A., Wilson, I. A., and Lerner, R. A. (1997). *Science* **278,** 2085–2092.

Blackwood, M. E., Jr., Rush, T. S., Romesberg, F., Schultz, P. G., and Spiro, T. G. (1998). *Biochemistry* **37,** 779–.

Braisted, A. C., and Schultz, P. G. (1990). *J. Am. Chem. Soc.* **112,** 7430–7431.

Buchbinder, J. L., Stephenson, R. C., Scanlan, R. C., and Fletterick, R. J. (1998). *J. Mol. Biol.* **282,** 1033–1041.

Charbonnier, J.-B., Carpenter, E., Gigant, B., Golinelli-Pimpaneau, B., Eshhar, Z., Green, B. S., and Knossow, M. (1995). *Proc. Natl. Acad. Sci. USA* **92,** 11721–11725.

Charbonnier, J.-B., Golinelli-Pimpaneau, B., Gigant, B., Tawfik, D. S., Chap, R., Schindler, D. G., Kim, S.-H., Green, B. S., Eshhar, Z., and Knossow, M. (1997). *Science* **275,** 1140–1142.

Chothia, C., and Lesk, A. M. (1987). *J. Mol. Biol.* **196,** 901–917.

Cochran, A. G., and Schultz, P. G. (1990). *Science* **249,** 781–783.

DelaFuente, G., and Sols, A. (1970). *Eur. J. Biochem.* **16,** 234–.

Doering, W. v. E., and Roth, W. R. (1962). *Tetrahedron* **18,** 67–74.

Facchiano, A. M., Colnna, G., and Ragone, R. (1998). *Protein Engineering* **11,** 753–758.

Fleming, I. (1978). *Frontier Orbitals and Organic Chemical Reactions,* New York, Wiley.

Foote, J., and Milstein, C. (1991). *Nature* **352,** 530–532.

Gershenson, A., Schauerte, J. A., Giver, L., and Arnold, F. H. (2000). *Biochemistry.* **39**(16), 4658–4665.

Gigant, B., Charbonnier, J.-B., Eshhar, Z., Green, B. S., and Knossow, M. (1997). *Proc. Natl. Acad. Sci. USA* **94,** 7857–7861.

Giver, L., Gershenson, A., Freskgard, P. O., and Arnold, F. H. (1998). *Proc. Natl. Acad. Sci. USA* **95**(22), 12809–12813.

Golinelli-Pimpaneau, B., Gigant, B., Bizebard, T., Navaza, J., Saludjian, P., Zemel, R., Tawfik, D. S., Eshhar, Z., Green, B. S., and Knossow, M. (1994). *Structure* **2,** 175–183.

Gray, J. V., Eren, D., and Knowles, J. R. (1990). *Biochemistry* **29,** 8872–.

Gruber, K., Zhou, B., Houk, K. N., Lerner, R. A., Shevlin, C. G., and Wilson, I. A. (1999). *Biochem.* **38,** 7062–7074.

Haynes, M. R., Stura, E. A., Hilvert, D., and Wilson, I. A. (1994a). *Proteins: Struct. Funct. Genet.* **18,** 198–200.

Haynes, M. R., Stura, E. A., Hilvert, D., and Wilson, I. A. (1994b). *Science* **263,** 646–652.

Heine, A., Stura, E. A., Yli-Kauhaluoma, J. T., Gao, C., Deng, Q., Beno, B. R., Houk, K. N., Janda, K. D., and Wilson, I. A. (1998). *Science* **279,** 1934–1940.

Hsieh-Wilson, L., Schultz, P. G., and Stevens, R. C. (1996). *Proc. Natl. Acad. Sci. USA* **93,** 5363–5367.

Hunter, C. A., Singh, J., and Thornton, J. M. (1991). *J. Mol. Biol.* **218**(4), 837–846.

Jaenicke, R., and Bohm, G. (1998). *Curr. Opin. Struct. Biol.* **8**(6), 738–748.

Kimura, M. (1982). *Molecular Evolution, Protein Polymorphism, and the Neutral Theory* Springer-Verlag, Berlin and Japan Scientific Societies Press, Tokyo.

Knowles, J. (1991). *Nature* **350,** 121–124.

Koshland, D. E. (1987). *Cold Spring Harbor Symp Quant. Biol.* **52,** 1–7.

Kristensen, O., Vassylyev, D. G., Tanaka, F., Morikawa, K., and Fujii, I. (1998). *J. Mol. Biol.* **281,** 501–511.

Lavallee, D. K. (1988). *Mol. Struct. Energy* **9,** 279–314.

Lesberg, C. A., Caruthers, J. M., Paschall, C. M., and Christianson, D. W. (1998). *Curr. Opin. Str. Biol.* **8,** 695–703.

Moore, J. C., Jin, H. M., Kuchner, O., and Arnold, F. H. (1997). *J. Mol. Biol.* **272**(3), 336–347.

Moore, J. C., and Arnold, F. H. (1996). *Nature Biotechnology* **14**(4), 458–467.

Mundorff, E. C., Hanson, M. A., Varvak, A., Ulrich, H., Schultz, P. G., and Stevens, R. C. (2000). *Biochemistry* **39**(4),627–632.

Nardini, M., and Dijkstra, B. W. (1999). *Curr. Opin. Struct. Biol.* **9,** 732–737.

Ollis, D. L., Cheah, E., Cygler, M., Dijkstra, B., Frolow, F., Franken, S. M., Harel, M., Remington, S. J., Silman, I., Schrag, J., Sussman, J. L., Verschueren, K. H. G., and Goldman, A. (1992). *Protein Engineering* **5**(3), 197–211.

Oue, S., Okamoto, A., Yano, T., and Kagamiyama, H. (1999). *J. Biol. Chem.* **274**(4), 2344–2349.

Paschall, C. M., Hasserodt, J., Jones, T., Lerner, R. A., Janda, K. D., and Christianson, D. W. (1999). *Angew. Chem. Int. Ed.* **38,** 1743–1747.

Patten, P. A., Gray, N. S., Yang, P. L., Marks, C. B., Wedemayer, G. J., Boniface, J. J., Stevens, R. C., and Schultz, P. G. (1996). *Science* **271,** 1086–1091.

Rajewsky, K. (1996). *Nature* **381**(6585), 751–758.

Romesberg, F. E., Santarsiero, B. D., Spiller, B., Yin, J., Barnes, D., Schultz, P. G., and Stevens, R. C. (1998a). *Biochem.* **37,** 14404–14409.

Romesberg, F. E., Spiller, B., Schultz, P. G., and Stevens, R. C. (1998b). *Science* **279,** 1929–1933.

Seidman, J. G., and Leder, P. (1978a). *Nature* **276**(5690), 790–795.

Seidman, J. G., Leder, A., Edgell, M. H., Polsky, F., Tilghman, S. M., Tiemeier, D. C., and Leder, P. (1978b). *Proc. Natl. Acad. Sci. USA* **75**(8), 3881–3885.

Serrano, L., Bycroft, M., and Fersht, A. R. (1991). *J. Mol. Biol.* **218**(2), 465–475.

Spiller, B., Gershenson, A., Arnold, F. H., and Stevens, R. C. (1999). *Proc. Natl. Acad. Sci. USA* **96**(22), 12305–12310.

Stemmer, W. P. C. (1994). *Nature* **370,** 389–391.

Sussman, J. L., Harel, M., Frolow, F., Oefner, C., Goldman, A., Toker, L., and Silman, I. (1991). *Science* **253**(5022), 872–879.

Thayer, M. M., E. H., Olender, Arvai, A. S., Koike, C. K., Canestrelli, I. L., Stewart, J. D., Benkovic, S. J., Getzoff, E. D., and Roberts, V. A. (1999). *J. Mol. Biol.* **291,** 329–345.

Tonegawa, S. (1983). *Nature* **302**(5909), 575–581.

Ulrich, H., Mundorff, E., Santarsiero, B. D., Driggers, E. M., Stevens, R. C., and Schultz, P. G. (1997). *Nature* **389,** 271–275.

Ulrich, H. D., and Schultz, P. G. (1998). *J. Mol. Biol.* **275,** 95–111.

Vogt, G., Woell, S., and Argos, P. (1997). *J. Mol. Biol.* **269,** 631–643.

Wedemayer, G. J., Wang, L. H., Patten, P. A., Schultz, P. G., and Stevens, R. C. (1997a). *J. Mol. Biol.* **268,** 390–400.

Wedemayer, G. J., Patten, P. A., Wang, L. H., Schultz, P. G., and Stevens, R. C. (1997b). *Science* **276,** 1665–1669.

Xie, S., and Lu, H. P. (1999). *J. Biol. Chem.* **274,** 15967–15970.

Yang, P. L., and Schultz, P. G. (1999). *J. Mol. Biol.* **294,** 1191–1201.

Zavodszky, P., Kardos, J., Svingor, A., and Petsko, G. A. (1998). *Proc. Natl. Acad. Sci. USA* **95**(13), 7406–7411.

Zhou, G. W., Guo, J., Huang, W., Fletterick, R. J., and Scanlan, T. S. (1994). *Science* **265,** 1059–1064.

This Page Intentionally Left Blank

MOLECULAR BREEDING: THE NATURAL APPROACH TO PROTEIN DESIGN

By JON E. NESS, STEPHEN B. DEL CARDAYRÉ, JEREMY MINSHULL, and WILLEM P. C. STEMMER

Maxygen, Redwood City, California 94063

The breeding of wild plants and animals into the agricultural domesticates we know today precedes Mendel's genetic studies by many millennia. Despite their lack of a formalized understanding, our ancestors harnessed the evolutionary power of sexual recombination of preexisting natural diversity to produce plants and animals with characteristics better suited to their needs. A recent advance in protein design, termed molecular breeding, allows protein engineers to homologously recombine multiple related genes by a process that closely mimics sexual recombination to generate functionally diverse libraries of chimeric proteins from which improved variants can be selected. Molecular breeding effects the permutation of diversity within a pool of related sequences and has proven to be an extraordinarily effective method to evolve proteins and pathways for better function.

I. INTRODUCTION

Just look around. Look at what nature has provided—things for basic necessities, such as the corn in breakfast cereal and cotton for clothing; things that bring pleasure, such as a pet dog or flowers. Most of these are not at all like what they were 5000 years ago. In fact, many have changed dramatically in only the last 100 years. This was accomplished by a process of selective breeding—a process as simple as mating one's two favorite dogs or saving the cream of the crop for next spring's

0065-3233/00 $35.00

planting. Of breeding, Darwin states: ''The key is man's power of accumulative selection: nature gives successive variations; man adds them up in certain directions useful to him'' (Darwin, 1859). Apart from the nature of selective pressure, Darwin recognized that the principles operating in the domestication and breeding of plants and animals were indistinguishable from those operating in natural selection. He recognized that natural variation within a population is the base material on which evolution and breeding depend. In addition to the similarity between breeding and natural selection, Darwin recognized that there must be a mechanism generating variation within the population from which nature or a breeder selects.

The block-wise exchange of homologous regions of chromosomes, which occurs during meiosis in sexually reproducing organisms (Roeder, 1997) and during genetic exchange in bacteria (Matic *et al.,* 1995; Ogunseitan, 1995), is by far the major generator of diversity. Preexisting diversity present within a population of organisms is shuffled to produce new variants. The fitter or more desirable organisms are those bearing a better mixture of beneficial alleles and fewer deleterious alleles; these are selectively propagated in nature or by a breeder. Accumulation of beneficial alleles over multiple generations under selective pressure, be it natural or imposed, can result in remarkable modifications from ancestral organisms. The role of homologous DNA recombination in accelerating evolution is evidenced by its chimeric signature in genes that have endured selective pressure. For example, the penicillin-binding proteins of β-lactam resistant *Neisseria* species show complex patterns of recombination between at least six ancestral species, differing in DNA sequence identity by up to 23% (Smith *et al.,* 1991). The diversity of life, both extinct and extant, and the rapidity with which microorganisms evolve (e.g., see van der Meer *et al.,* 1992) are testaments to the effective mechanisms nature has developed and breeders have tapped to improve useful plants and animals (Burbank *et al.,* 1914). Recognizing the power of selective breeding, protein engineers have captured sexual recombination in the test tube to rapidly improve the products of genes. The highly active, functionally diverse gene libraries generated by molecular breeding[1] have extended directed evolution to a plethora of proteins for which only limited throughput screens are feasible (Chang *et al.,* 1999; Christians *et al.,* 1999; Minshull and Stemmer 1999; Ness *et al.,*

[1] Maxygen is the leader in the emerging field of directed molecular evolution, the process by which novel genes are generated for commercial and research purposes. Maxygen's proprietary technologies, known as MolecularBreeding™, mimic the natural process of evolution and bring together advances in classical breeding, molecular biology, and genomics.

1999; Patten *et al.,* 1997; Zhang *et al.,* 1997). Molecular breeding has also been extended beyond proteins to pathways and viruses, as well as partial genomes (Crameri *et al.,* 1997; Soong *et al.,* 1999).

II. The Need for Better Proteins

Our ancestors learned to exploit a variety of natural enzymatic transformations for the production of food and drink, such as the malting of barley and its fermentation to beer. Today enzymes have found many applications. Examples include the multi-enzymatic production of high fructose corn syrup (Bhosale *et al.,* 1996; Hagedorn and Kaphammer, 1994), the synthesis of acrylamide (Kobayashi *et al.,* 1992), and the use of streptokinase (Wong *et al.,* 1994; Wu *et al.,* 1998) for the therapeutic degradation of arterial blood clots. There are many benefits from using enzymes. In contrast to chemical catalysts, enzymes are nontoxic, biodegradable, and can be produced by fermentations of cheap, renewable feedstocks. In chemical synthesis, the exquisite specificity (substrate-, regio-, and stereo-selectivity) and efficiency of enzymes can obviate the need for protecting and deprotecting groups and simplify end-product purification. In addition, enzymatic processes generally require more moderate reaction conditions and simpler, more flexible production units. The use of biocatalysts (both enzymes and whole cells) as benign alternatives to chemical catalysts for producing renewable chemicals, pharmaceuticals, polymers, and fuels is integral to realizing current visions of sustainable development (Nedwin, 1997).

Despite these advantages, there are discrete performance limitations that have impeded biocatalysts from realizing their industrial potential. These limitations originate in the natural physiological role that biocatalysts play. Most enzymes have evolved to function optimally on a narrow set of substrates and under the precise conditions (temperature, ionic strength, pH, etc.) of their natural niche. The efficiency of enzymes allows for their transient use and for a cell to recycle their amino-acid components, thus wild-type enzymes often are unstable and prone to chemical and biological degradation. Although these characteristics provide a benefit to a cell in nature, they are suboptimal for the biotechnologist hoping to exploit a biocatalyst. The substitution of an enzyme for a chemical process often requires that the protein is active for many hours, capable of functioning under reaction conditions alien to its natural milieu (such as extremes of temperatures and pH or in the presence of organic solvents), able to use nonnatural substrates, and cost effective relative to alternative chemical catalysts.

Proteins are also used clinically to treat a variety of diseases. Erythropoietin stimulates erythrocyte production in kidney dialysis and chemotherapy patients. Granulocyte stimulating factor enhances immune systems compromised by cancer treatments. Cytokines such as interferons and interleukins are used for their anti-viral and anti-tumor activities. Other proteins are used to inhibit or stimulate blood clotting. For the most part, the pharmaceutical protein industry relies on cloning native human genes and expressing and purifying their products in recombinant form.

As with proteins used as biocatalysts, natural human proteins may require optimization for use as pharmaceuticals. One frequent limitation is protein half-life. Another is selectivity. Cytokines often bind to multiple receptors on different cell types, with different physiological responses resulting from the different binding events. Thus, while a therapeutically relevant dose of a protein pharmaceutical may activate a targeted biological pathway, it may also increase harmful or even fatal side effects due to activation of secondary pathways. In this case, a cytokine that activated only one of the pathways would clearly be desirable. Another limitation of naturally occurring proteins is that they often require high doses (milligrams to grams), making their use as therapeutics economically impractical. Yield of active recombinant protein production in a heterologous system is another property that could be improved, as was done for the expression of the jellyfish green fluorescent protein (GFP) in *Escherichia coli* (Crameri *et al.,* 1996). The strategies described below are most frequently related to enzyme activity, yet they clearly also apply to modifying the properties of interest in proteins for pharmaceutical and a plethora of other uses.

III. Strategies for Optimizing Proteins

Natural selection and classical breeding both employ the same empirical strategy of creating variants and selecting those that perform best—that is the essence of all protein tailoring methods. They differ only in the sources of sequence variation and the methods by which this diversity is tested.

A. Natural Diversity

The most reasonable source of starting points is nature. If one protein does not perform exactly as required, perhaps a homolog isolated from a different organism will. For example, an enzyme that functions well in high salt conditions may best be isolated from a halophile. On the other hand, properties such as lack of product inhibition or ability to function in an organic solvent may never be found by searching through nature, simply because a cell's survival has never depended on these characteristics. Naturally occurring proteins can be thought of as analogous to the wild ancestors of modern crops or domesticated animals: good starting points, but unlikely to possess the full range of properties required for human purposes (Diamond, 1997). A different source of variants is to take an available protein whose function most closely approximates that desired, and make and test variations on that sequence.

B. Variations on a Single Sequence

1. Rational Design

Structure-based protein design relies on knowing the structure of a protein and on tools for molecular modeling to predict favorable amino acid changes. A small number of promising modifications are introduced into the gene by methods such as oligonucleotide-directed mutagenesis (Kunkel *et al.,* 1991), and the effects tested. This method tests only a small number of variations. However, our understanding of protein function, structure, folding, and interaction is still sufficiently imprecise that no matter how good the available structural data, the variants frequently fail to show the desired improvements. Although the approach has proven fruitful and general strategies for tailoring a few properties are slowly being illuminated, the approach is impeded by assumptions that discount the complexity of biological systems (Rubingh, 1997). Although recent advances for visualizing proteins as the dynamic structures they are (Arnold and Ornstein, 1997) and for the *de novo* design of protein folds (Dahiyat and Mayo, 1997) hint at a future when proteins may be built to desired specifications, the prospect of designing proteins for specific functions is still a long way off.

2. Random Point Mutations

Iterated random point mutagenesis coupled with a screen or selection to evolve an improved protein (Arnold, 1998a; Shao and Arnold, 1996) is a strategy for molecular evolution that stems from classical strain improvement of industrial microorganisms (described below). Random point mutagenesis of the gene for the protein to be improved is typically performed by error-prone PCR (Arnold, 1998a; Cadwell and Joyce, 1992; Cadwell and Joyce, 1994; Chen and Arnold, 1991; You and Arnold, 1996), but exposure to chemical mutagens (Taguchi *et al.,* 1998), or mutator strains (Bornscheuer *et al.,* 1998; Low *et al.,* 1996) have also been used. Because most mutations are detrimental or neutral (Shafikhani *et al.,* 1997; Suzuki *et al.,* 1996), a low mutation frequency is employed to generate approximately one to two amino acid changes per protein. A sample of the library is subjected to an appropriate selection or screen to identify those variants that have the desired improvements. The process is repeated with the single best performer, often with increasingly stringent selection pressure or screening criteria. The result is the ''asexual'' evolution of an improved protein by stepwise accumulation of single mutations. The advantage of this method, in contrast to structure-based engineering or random cassette mutagenesis, is that little or no information regarding protein structure and function is required, and few assumptions are made. Further, the process generally provides new protein sequences with measurable improvements in a desired activity in relatively small libraries ($\sim 10^4$/iteration). Amino acid changes contributing to an evolved phenotype are frequently scattered throughout the protein sequence; changes whose effects could not have been accurately predicted or calculated (Arnold, 1998a; Spiller *et al.,* 1999). The unbiased nature of the process affects diverse aspects of protein activity that are not easily modeled, such as transcription, translation, protein folding, and protein-protein interaction with host systems. Individual clones commonly have improvements in several of these properties. Despite the success of iterated mutagenesis compared to rational approaches, the process is still impeded by the low quality of random mutations and the inherent limitations of an asexual evolutionary process that accumulates mutations at a rate of only about one per cycle (discussed below).

3. Oligonucleotide Randomization

A feature of random point mutagenesis by the methods described above is that it is limited to single point mutations. Only one nucleotide in any codon is usually changed, which means that on average less than six amino acids may be accessed (Jespers *et al.,* 1997) (the remaining fourteen amino acids could only be accessed by double- or triple-base mutations). Because the amino acids that are accessible with a single base mutation tend to be conservative replacements for the original residue, this important bias results in mutations that are more conservative than fully random. However, saturation mutagenesis is a way to circumvent this bias, for example, by using a cassette containing a randomized codon to replace the native codon with an equal mixture of all possible codons (Reidhaar-Olson *et al.,* 1991). Because this approach is the most mutagenic, it can only be applied to a small part of a protein and is used primarily when enough structural information is available to target a certain area, but not enough for designing and testing specific single mutations. This approach therefore lies between the purely structural-based and random approaches. Molecular modeling and other rational approaches are employed to identify regions of a gene to target by random mutagenesis. Screening the mutant clones for variants of desired function identifies random mutational solutions lying within the targeted region. This approach has found success (Black *et al.,* 1996), but remains limited by several factors. These include its imposition of rational criteria, the ability to identify critical, contiguous regions for mutagenesis, the fact that beneficial mutations are often found throughout a polypeptide, and the low quality of random mutations that requires the testing of large libraries when multiple random changes are introduced into a critical region, such as the substrate binding pocket of an enzyme. The latter limitation can be reduced by using a biased randomization, for example, 70% of the wild-type base and 10% of each of the other bases (Reidhaar-Olson *et al.,* 1991) or doping for desired subsets of amino acids (Arkin and Youvan, 1992). Variations of this method include scanning saturation mutagenesis (Chen *et al.,* 1999) and recursive ensemble mutagenenesis (Delagrave *et al.,* 1993; Delagrave and Youvan, 1993). These processes all increase the genetic variation accessible at a codon (or contiguous series of codons); however, the methods require construction of many libraries for complete scanning of a protein and are difficult to iterate.

C. Recombination of Diversity

A fundamental difference between the mutagenesis methodologies previously described and the examples of natural evolution and breeding described in the Introduction (Section I) is the role played by recombination. Homologous recombination has two major effects on an evolving system. First, it avoids reinventing the wheel: Once a mutation that confers an improved phenotype has entered a population, recombination allows it to be tested in combination with all other beneficial mutations, rather than all of these combinations having to be derived *de novo.* Second, it allows efficient removal of deleterious mutations by replacing them with wild-type sequence, thereby avoiding the downward spiral of fitness resulting from deleterious mutations in asexual populations (Muller's ratchet) (Muller, 1964). Not only is recombination a feature of all biological systems, but computational simulations of recombination show dramatic increases in the speed and range of evolution when recombination is included in the algorithms (Forrest, 1993; Gibson, 1989; Holland, 1975; Kelly, 1994). Molecular breeding (also called DNA shuffling) was developed to mimic this essential feature of natural evolution. The sexual process of molecular breeding (Stemmer, 1994a; Stemmer, 1994b) has supplanted iterated random mutagenesis as the most efficient and rapid method for directing the evolution of nucleic acids and proteins.

The most widely used format for molecular breeding is *in vitro* fragmentation and reassembly of DNA (Fig. 1). In this format, DNA from a pool of selected mutants is randomly fragmented (e.g., with DNase I) and reassembled in a primerless DNA amplification reaction. The reassembly reaction is recombinogenic because fragments from one DNA sequence can prime homologous regions of different DNA sequences by template switching. In addition, the level of mutagenesis can be adjusted with the appropriate choice of DNA polymerase and reaction conditions (Zhao and Arnold, 1997c).

1. *Recombination of Single Sequences by DNA Shuffling*

The power of the combinatorial nature of DNA shuffling was first demonstrated using the TEM-1 β-lactamase (Stemmer, 1994b). When expressed in *E. coli,* the TEM 1 β-lactamase provides low-level resistance to the poorly hydrolyzed antibiotic cefotaxime. Three cycles of DNA shuffling (error-prone, to create the initial diversity), selection, and pooling of the most cefotaxime resistant mutants improved resistance from a minimal inhibitory concentration (MIC) of 0.02 μg/ml to 320 μg/ml for a 16,000-fold increase in resistance. Two rounds of backcrossing (i.e., shuffling with a molar excess of the parental gene) resulted in a mutant that was 32,000-fold more drug resistant. Backcrossing (Stemmer, 1994b) of genes that encode improved variants is a method to identify beneficial mutations while flushing out neutral and deleterious mutations (Zhao and Arnold, 1997b). In addition to improved turnover, this mutant retained a promoter mutation that resulted in a twofold increase in expression over the parental enzyme, illustrating the ability of DNA shuffling to solve a problem through more than one route.

This example demonstrates both of the recombination-derived advantages of DNA shuffling. The first advantage is that DNA shuffling can create combinations of distant, separately selected residues. In a previous approach to increase cefotaxime resistance by molecular modeling of the known structure of the TEM-1 β-lactamase, three active site loops were randomized separately (Palzkill and Botstein, 1992). These separate libraries yielded the key E104K and G238S mutations, resulting in four- and eight-fold increases in cefotaxime resistance, respectively. In combi-

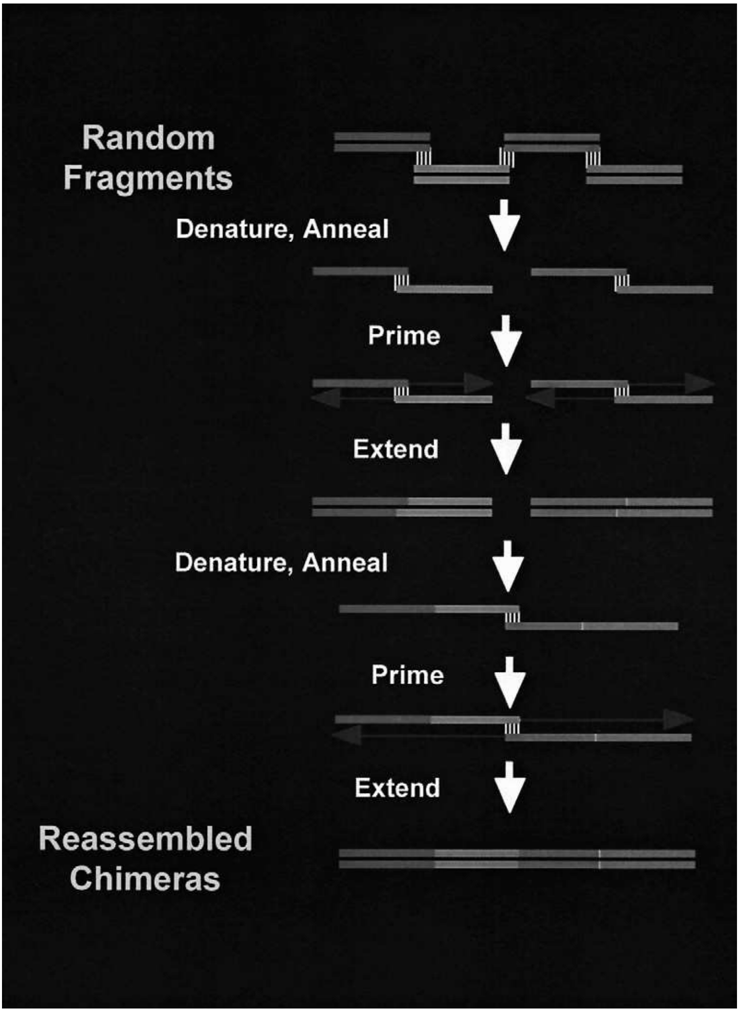


FIG. 1. DNA shuffling by fragmentation and reassembly. A pool of DNA sequences is randomly fragmented (e.g., by treatment with DNase I). The gene fragments are assembled into a library of full-length chimeric genes by repeated cycles of denaturation, annealing, and DNA polymerase extension.

nation, however, these two changes result in a 500-fold increase. Although shuffling quickly yielded this double mutation (in addition to other mutations that provided a further sixty-four-fold improvement), the libraries constructed in separate loops based on rational design could not yield such a combination mutant. The second advantage of recombination is the purging of the excess of deleterious mutations, which tend to mask the effect of the beneficial mutations. A control experiment using three cycles of error-prone PCR resulted in only a sixteen-fold improvement. Although in principle this method should have yielded the combination of the two key mutations, and the 500-fold increase in MIC, the combination was not seen in practice, probably as a result of the much higher rate of deleterious mutations. Reversal of such deleterious mutations by replacement with wild-type sequence via recombination is efficient, but reversal of deleterious mutations by error-prone PCR is very inefficient. This second advantage is also demonstrated in the example of GFP evolution by DNA shuffling (Crameri *et al.,* 1996). All three of the mutations that resulted in a 45-fold increase in fluorescence were present in a single mutant following the first round of mutagenesis. The effect of two subsequent rounds of shuffling was simply to remove three additional mutations by recombination with wild-type sequences within the pool.

The ability of DNA shuffling to generate a large number of new combinations of beneficial mutations that originated on separate DNA molecules is the key accelerator of directed evolution, since the mutations are often additive or even synergistic in effect (Chen and Arnold, 1991; Matsumura *et al.,* 1986; Shaw *et al.,* 1999; Wells, 1990). By contrast, in nonrecombination approaches the single best parent is typically selected as the sole parent to create the mutant library (either by random or rational mutagenesis) that is to be screened in the next cycle; this effectively discards all except the single best mutation, as well as combinations of mutations that were painstakingly identified in the previous cycle. Iterated cycles of DNA shuffling of a single starting gene using random point mutations as a source of diversity have been successively used to evolve proteins with enhanced activity (Stemmer, 1994b), altered substrate specificity (Zhang *et al.,* 1997), improved protein folding (Crameri *et al.,* 1996), thermostability (Giver *et al.,* 1998), solvent tolerance (Moore *et al.,* 1997), and resistance to chemical modification (Matsumura *et al.,* 1999). Table I summarizes a selection of published examples of DNA shuffling of a single starting sequence.

Although DNase I fragmentation and reassembly is the most widely used format for DNA shuffling, a number of alternative methods have been developed. All rely on the same underlying principle that the most efficient way to explore all possible combinations and permutations of sequences is by recombination. Two proved alternative methods of DNA shuffling are the staggered extension process or StEP (Zhao *et al.,* 1998) and *in vivo* shuffling in *Saccharomyces cervisiae* (Cherry *et al.,* 1999). Shuf-

TABLE I

Systems Improved by DNA Shuffling of a Single Starting Sequence

System	Comments	Reference
Single Proteins		
TEM-1 β-lactamase	3 cycles of shuffling and 2 cycles of backcrossing, 32,000-fold increase in antibiotic resistance	Stemmer, 1994
β-galactosidase	7 cycles, 66-fold increase in fucosidase specific activity, 1000-fold increase in substrate specificity	Zhang *et al.*, 1997
Green fluorescence protein	3 cycles, 45-fold improvement in fluorescence as a result of improved protein folding	Crameri *et al.*, 1996
Human antibody	8 cycles of shuffling and 2 cycles of backcrossing, >440-fold increase in avidity	Crameri *et al.*, 1996
Mouse antibody	100-fold increase in expression level	Crameri *et al.*, 1996
Pathways		
Arsenate degradation pathway	3 cycles, 40-fold improvement in arsenate resistance	Crameri *et al.*, 1997

fling by random primers (Shao *et al.,* 1998) is not currently widely practiced.

StEP is a PCR-like reaction consisting of a mixture of full-length templates with different beneficial mutations and flanking oligonucleotide primers. In contrast to PCR, StEP employs an extremely abbreviated polymerization step to generate partially extended fragments, which undergo template switching during subsequent cycles of the fragment reassembly reaction. Although StEP can be carried out in a single tube, crossover frequencies are limited by the rapidity with which cycles can be performed.

In vivo DNA shuffling in *S. cerevisiae* uses the cell's highly efficient double-strand DNA break repair pathway to obtain recombination. Yeast cells are co-transformed with a linearized vector and a series of overlapping DNA fragments (e.g., restriction fragments) that together comprise the target sequence. Vector replication requires recircularization of the fragments and vector by a series of *in vivo* recombination events in the homologous overlapping regions. As little as 15 to 30 bp of contiguous identity is all that is required for recombination in yeast (Manivasakam *et al.,* 1995).

Although we have developed and are continuing to develop a range of alternative recombination formats (*in vitro, in vivo,* and combination methods), the *in vitro* methods such as fragmentation and reassembly are currently preferred for most applications due to their versatility and control.

2. Recombination of Multiple Homologs from Natural Diversity by DNA Shuffling

The diversity of sequence variations that exists in natural populations of organisms is likely to be very old and highly stable. It is a reservoir of diversity that has proved itself functional and useful. Although this ''proved'' diversity originated from random mutagenesis, it is much more conservative. For example, introduction of random single amino-acid mutations into a typical protein leads to a high level of nonfunctional ''knockout'' mutants, yet in dog breeding where one is shuffling two dog genomes that can differ by over a million different single base mutations, most of the puppies that are born are fully functional, suggesting high library quality. Such a result could only be obtained if this large pool of natural diversity was extremely conservative. The fact that such a large fraction of the new combinations of this diversity are so well tolerated suggests that perhaps many of the mutations have already been selected for function and permutability. By contrast, ten generations of randomly mutating dogs will produce a very sick dog rather than a useful new breed.

On the level of the individual gene as well as of the entire organism, changes have been accumulated over millions of years. In the example given above, only genes that function within the context of an entire functional dog have persisted. Similarly, for a single gene, nature has only maintained sequences that are functional within the sequence and structural context of the entire protein and within the complex environment of the whole cell. However, together with the sequence divergence, there has been divergence in a range of properties of gene families. For example, enzymes have evolved to function in diverse physical and chemical conditions (Narinx *et al.,* 1997), to accept new substrates (Scanlan and Reid, 1995), or even to perform fundamentally different chemical reactions (Babbitt *et al.,* 1995). Receptors and ligands have diverged as they co-evolved, maintaining tight binding with each other, but greatly reduced binding to the receptor ligand pairs that have co-evolved in other species. Consequently, as with dog breeding, exchanging blocks of these sequences results in a library containing a large proportion of active members, with a high degree of phenotypic diversity. The permutation of natural sequence diversity encoding amino-acid changes and neutral mutations, as well as deletions and insertions, produces libraries of progeny that are quite different in sequence and in their combinations of characteristics from any of the parents and therefore represents a broad but sparse sampling of sequence space. Sparse sampling is obtained because the exchange of large sequence blocks by shuffling means that neighboring mutants generally differ at multiple amino-acid positions (Fig. 2). Just as the recombination of dog genomes resulting from sexual reproduction produces functional offspring that differ from either parent, so the molecular breeding of the genes encoding a closely related family of proteins results in a library of functional but different molecules. However, unlike classical breeding, molecular breeding is not limited to two parents and thus bypasses natural species barriers. In addition, molecular breeding is fast with a cycle time of days rather than months or years required for a cycle of classical breeding (Table II).

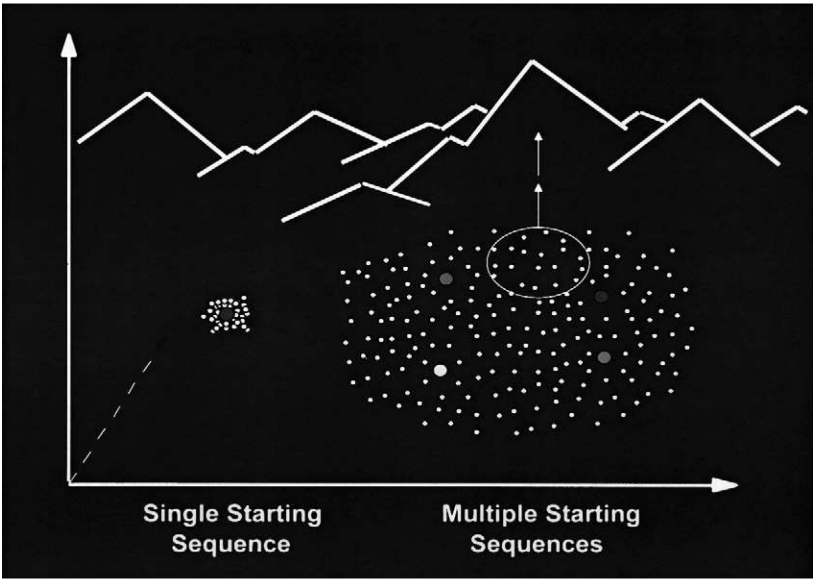


FIG. 2. Searching sequence space by molecular breeding versus random mutagenesis of a single starting sequence. Random mutagenesis yields clones with a few point mutations. The approach is suitable for "hill climbing" to a local performance maximum. Family shuffling yields chimeras that typically have many changes relative to parental sequences and other progeny. At equal library size, the increased sequence diversity results in a sparse sampling of a much greater area of sequence space, allowing much more promising regions to be found and subsequently explored at increased sampling density (Crameri *et al.,* 1998).

TABLE II
Classical Breeding versus Molecular Breeding

Classical breeding	Molecular breeding
Cycle time = years	Cycle time = days
Whole genome	Genes, pathways, genomes
Breed within species	Breed across species
Two parents	One to many parents
Limited control	Multilevel control
Complex selection pressure	Focused selection pressure
Whole plants and animals	Applicable to microbes, cells, whole organisms

The first published example of molecular breeding of natural diversity involved the recombination of four *ampC* β-lactamases that shared 58% to 82% amino-acid identity (Crameri *et al.,* 1998). A library of recombinants was tested for ability to confer increased bacterial resistance to moxalactam, an antibiotic that is poorly degraded by *ampC* β-lactamases. Screening 50,000 members of this library produced a variant that was different from its closest parent at over 25% of its amino acids and conferred 270-fold greater resistance to moxalactam than did the best parent. This compares with an up to eightfold improvement found by single sequence DNA shuffling of any of the four parent genes separately (Fig. 3, see color insert). This enzyme also conferred resistance to a number of other β-lactams.

The shuffling of twenty-six subtilisins and over twenty human interferons are two recent examples that demonstrate the power of molecular breeding to generate high quality and functionally diverse libraries of useful proteins that serve as rich sources of hits when screened for desired properties (Chang *et al.,* 1999; Ness *et al.,* 1999).

a. Interferons. The evolution of pharmaceutical proteins by molecular breeding has recently been demonstrated by Chang *et al.* (Chang *et al.,* 1999). They built a high quality α-interferon library by shuffling more than twenty human α-interferons that shared nucleotide homologies of 85% to 95% (much greater than in the β-lactamase example previously cited). The library was screened as pools of clones for increased antiviral activity, measured by the protection of murine cells against a challenge with lymphocytic choriomeningitis virus, and positive pools were then deconvoluted. Using a total of sixty-eight assays to screen a library of 1672 previously unscreened clones from a single round of recombination, a variant was identified with 135,000-fold greater specific activity than the parent interferon, Hu-IFN-α2a. A second round of shuffling yielded a variant with a 285,000-fold higher specific activity than Hu-IFN-α2a, 185-fold higher specific activity than Hu-IFN-α1, and four-fold more active than the most potent murine interferon, Mu-IFN-α4, despite the fact that the human and murine sequences are only about 65% identical (Fig. 4, see color insert). Importantly, the best three clones were all composed of multiple segments from known human interferons, and contained no new point mutations, suggesting that these clones are much less likely to be immunogenic.

Mice and humans (and their respective cytokines and receptors) have been evolutionarily separated for over 100 million years. Consequently, the murine interferons differ from all of the human interferons at fifty-six to seventy-two amino-acid positions. Nevertheless, it was possible, simply by recombining human sequences, to produce an interferon with an activity greater than that of the natural mouse protein. Sequences that had been ''pre-tested'' in humans to function as α-interferons contained the information necessary to build a protein that functions well with murine cells. Even though it was not possible to reconstruct the exact sequence of murine α-interferon from the human genes, it was possible to reconstruct its function.

b. Subtilisins. A second example of molecular breeding of a large family of proteins is that of subtilisin. Subtilisins are commercially important serine endoproteases, valued for a range of applications, perhaps most notably as additives to laundry detergents for hydrolysis and solubilization of protein stains (Bott and Betzel, 1996). With annual sales of about $500 million, it is not surprising that subtilisin is one of the best understood proteins and a frequent target for improvement using both structure-based design and random mutagenesis (Ballinger *et al.,* 1996; Bryan *et al.,* 1986; Graycar *et al.,* 1999; Kano *et al.,* 1997; Russel and Fersht, 1987; Wells and Estell, 1988; You and Arnold, 1996). As with most industrial enzymes, incremental improvements in performance are significant. A major challenge in the rational design or directed evolution of industrial enzymes is that performance is not defined by any single property, but by a complex mix of parameters. Although rational design and random mutagenesis can improve single properties, such as thermostability or activity in organic solvent, it is often at the expense of other critical properties (Patkar *et al.,* 1998; Shoichet *et al.,* 1995), making it difficult to obtain an enzyme that is optimized for several of the important performance criteria. Just as multiple traits in plants and animals can be recombined by classical breeding, multiple enzyme properties can be recombined by molecular breeding. Ness *et al.* (Ness *et al.,* 1999) demonstrated this by using DNA shuffling to breed twenty-five subtilisin

Legends for Color Insert

FIG. 3. (a) Comparison of single sequence shuffling and family shuffling of cephalosporinase. (b) Computer model of winning chimera created from the known structure of the *Enterobacter cloacae* protein (Crameri *et al.,* 1998; Lobkovsky *et al.,* 1993). The predicted structure of the α-chain backbone is within an r.m.s deviation of O.766Å from the known structure. The segments derived from *Enterobacter* are shown in blue, those from *Klebsiella* are shown in yellow, and those from *Citrobacter* are shown in green. The thirty-three amino acid point mutations are shown in red. The enzyme differs by 102 amino acids from the *Citrobacter* enzyme, by 142 amino acids from the *Enterobacter* enzyme, by 181 amino acids from the *Klebsiella* enzyme, and by 196 amino acids from the *Yersinia* enzyme.

FIG. 4. Summary of antiviral activities of native and evolved IFN-αs. The antiviral activities of purified protein for native Mu-IFN-αs and Hu-IFN-αs as well as evolved IFN-αs on murine L929 cells are shown. One unit of activity corresponds to half-maximal protection from lethal ECMV viral challenge. Arrows on the right indicate fold improvement of the winning IFN-α (IFN-α-CH2.1) relative to Hu-IFN-α1 and Hu-IFN-α2a (Chang *et al.,* 1999).

FIG. 5. Activities of 654 active clones from the shuffled subtilisin library compared to twenty-six parents. Relative activities of each clone in five screens are plotted as concentric circles. Each color represents one of the five screening conditions: pH5.5 (orange), pH7.5 (blue), pH10 (dark red), thermostability (yellow), and activity in 35% DMF at pH 7.5 (green). The area of the circle is proportional to the activity in the five assays relative to the best parent in each assay.

gene fragments obtained from a panel of mesophilic *Bacillus* isolates with the full-length gene for Savinase, a leading industrial protease (Graycar *et al.,* 1992; Hastrup *et al.,* 1989). The diversity of subtilisins used was much greater than that used in the interferon and β-lactamase examples. Pairwise identities of the DNA sequences were as low as 56.4% (protein sequences homology as low as 63.7%). A small library of 654 active clones was screened for thermostability, solvent stability, and pH dependence (at pH5, pH7.5, and pH10), three properties that are of commercial importance for subtilisin and of general concern for other industrial enzymes and biocatalysts.

The vast array of functional diversity generated in this experiment is shown in Figure 5 (see color insert). The frequency of improved clones ranged from 4% to 12% of the active library in any single parameter. In addition, the diversity of combinations of properties ranged well beyond the properties of the parental enzymes. Sequence analysis of some of the best performing clones under each set of conditions revealed that variants with similar properties could be encoded by very different sequences. Thermostability, for example, could be conferred by any one of at least three different genetic elements. In many applications, a family of natural sequences is known, but for historical reasons all of the characterization has focused on only one of the family members, typically one whose structure has been determined. Molecular breeding allows less-well characterized sequence homologs (and even partial or inactive sequences) to be incorporated into the breeding pool in any molar ratio desired. The screening of pluripotent enzyme libraries generated by molecular breeding of a handful of homologs provides an economical alternative to rational design or bioprospecting for leads that meet the multiple parameters required for commercialization. In addition, the ability of molecular breeding to demarcate functional sequence elements is likely to be a valuable tool for building structure-function databases for guiding protein design in the future. Table III summarizes a selection of published examples of molecular breeding of multiple related sequences.

IV. Screening Is Key

As Darwin observed, the only difference between the breeding of domestic plants and animals and the evolution of wild organisms is how the selection is applied. In nature, adaptation occurs in response to the environment: Diverse ecological niches give rise to a diversity of organisms to exploit them. Organisms undergo a very low level of random mutagenesis and those mutations that confer a competitive advantage (such as the ability to utilize a new nutrient source, survive at a higher or lower temperature, or kill a neighbor) are maintained in organisms that consequently grow and colonize a new niche.

Table III
Systems Improved by Molecular Breeding of Homologous Sequences

System	Comments	Reference
Cephalosporinase	4 sequences, 58%–82% DNA identity, 1 round, 270-540-fold increase in antibiotic resistance	Crameri, Raillard *et al.,* 1998
Thymidine kinase	2 sequences, 78% DNA identity, 4 rounds, 32-16,000-fold decrease in levels of AZT required to sensitize *Escherichia coli*	Christians, Scapozza *et al.,* 1999
α-interferon	>20 sequences, 85%–95% DNA identity, 2 rounds, 185–285,000-fold improvement in specific activity	Chang, Chen *et al.,* 1999
Subtilisin	26 sequences, 56%–99% DNA identity, 1 round, up to 4-fold improvement in 5 properties	Ness, Welch *et al.,* 1999

In classical breeding, the criteria for survival are altered to favor human needs and many of the restraints of natural selection are removed. For example, with people to protect them from predators, cattle are no longer subject to all the rigors of competing in the wild. Humans select and breed from those animals with the highest milk and meat productivity, while alertness and aggression are not only no longer required, but are in fact characteristics that detract from the ''performance'' of a large domesticated mammal. In a similar way, enzymes used in industrial processes need not be constrained by what is useful to an organism in the wild. For example, a property like product inhibition, which is a critical function of cellular economics, is no longer desirable. The protein economics attempts to subvert the function of the enzyme to be maximally productive and stable under conditions dictated by bioprocess engineers. Like breeders, protein engineers select those enzymes that perform best under the desired conditions.

The challenge of protein design by molecular breeding is the formulation of a screen that precisely emulates the final process conditions. This can be difficult to do in a high throughput format. A powerful approach is the employment of multitiered screens that sample decreasing numbers of clones with increasing scrutiny, ultimately ending with a handful of variants that are tested in the final process. Variants with improved performance under process conditions are carried forward to the next cycle of alteration and screening.

All evolutionary screens require some way to link phenotype with genotype. Recent technologies for linking genotype and phenotype have expanded the accessible library size by many orders of magnitude. For example, cell surface display (Daugherty *et al.,* 1998; Daugherty *et al.,* 1999; Georgiou *et al.,* 1997), phage and virus display (Hodits *et al.,* 1995; Smith *et al.,* 1998; Winter *et al.,* 1994), and ribosome display (Hanes *et al.,* 1999; Hanes *et al.,* 1998; Hanes and Pluckthun, 1997; Mattheakis *et al.,* 1994; Roberts and Szostak, 1997) provide access to libraries of 10^9–10^{14} variants. Unfortunately, screening these libraries remains limited to primarily affinity enrichment. For this reason applications have generally been limited to the identification of polypeptides (mostly antibodies and peptides) that bind tightly to a desired ligand. Although creative exceptions have been reported (Baca *et al.,* 1997; Janda *et al.,* 1994; Smiley and Benkovic, 1994), catalysis or intracellular function have not been conveniently addressed, nor has the technology realized general application such as screening for catalysis.

In some situations it is possible to develop a way to use genetic selection to identify mutants (Black and Loeb, 1993; Naki *et al.,* 1998). By coupling gene function to cell survival (e.g., the acquisition of an essential nutrient, the destruction of a toxic compound, or the ability to activate or complement an existing host metabolic pathway), up to about 10^{12} variants (using combinatorial infection) can be tested. However, the approach is limited by the fact that cells under selective pressure often find unexpected ways to grow (e.g., via genetic reversion or activation of cryptic functions). In addition, selections often have a limited dynamic range. Moreover, it is difficult to distinguish between the specific activity of an enzyme and an increase in its expression level. In general, one must evaluate promising variants from several different angles to avoid undesirable solutions, and must confirm that a phenotype is linked to the gene of interest (i.e., that plasmids from survivors confer viability to a naive cell). Perhaps the main limitation is that a genetic selection for a particular problem is not always obvious. When available, selections are most useful as the first tier (''filter'') of a multitiered screening program, which needs to be followed by subsequent, lower throughput but higher veracity screens to evaluate the positive clones.

Fluorescence-based cell sorting also allows screening of large numbers of variants (10^5–10^9). In general, this requires that a fluorescent product is formed and then retained within the cell. An elegant example is the evolution of a P450 enzyme to use hydrogen peroxide in place of the normal NADH cofactor in the hydroxylation of aromatic substrates (Joo *et al.,* 1999). Cells containing active horseradish peroxidase were transformed with a P450 library. Hydroxylated aromatic compounds were linked by the peroxidase to form fluorescent compounds that could then be detected by FACS or digital imaging. The intensity of the fluorescence increased with activity of the P450. In addition, different hydroxylation products resulted in different fluorescence spectra, so that an indication of regioselectivity could be obtained.

Agar plate screens allow the rapid analysis of up to 10^6 different colonies. Screenable phenotypes include enzymes that give rise to a color or fluorescence change in a diffusible substrate (Yang, 1994) or that form a halo around a producing cell because of the degradation of a insoluble substrate, such as the proteolysis of casein in agar plates containing skim milk (Cunningham and Wells, 1987). Plate-based bioassays can be used to detect and quantitate the production of a toxic compound (such as an antibiotic) as a zone of killing of an overlaid tester strain around the producing colony. Coupling of an agar plate-based screen with automated colony picking of positive clones provides a powerful first screen in a multi-tiered screening approach (Ness *et al.,* 1999).

In general, assays that quickly analyze an enormous population of variants tend to compromise the characteristics for which the variants are screened, sacrificing veracity and appropriateness for increased throughput. Consequently, while a high capacity assay may be an efficient way to reduce the number of candidates, it is essential to progress to a more accurate screen before doing additional cycles of recombination. Conditions can typically be manipulated in microtitre plates to give a reasonable approximation of the final process conditions. With a robotic system and a simple homogeneous fluorogenic or chromogenic assay, it is possible to test up to 10^6 variants in microtiter plates. In cases demanding a more complicated screen, such as an assay for stereoselectivity (Janes and Kazlauskas, 1997), about 10^4 clones can be screened in microtiter plates. Finally, manufacturing conditions can be even better simulated in flasks, fermentors, or small reactors, using the protein in its final whole-cell or purified protein form. Low throughput but accurate physical methods such as HPLC, mass spectroscopy, and gas chromatography measure catalytic activity accurately. These high veracity methods are useful as the final assay in a round of directed evolution to ensure that the positive variants really have the required activities and properties.

When optimizing a single property in model systems, such as thermostability or tolerance to an organic solvent, simple high throughput screens have proved adequate (Zhao and Arnold, 1997a). However, to address more complex problems, such as the generation of an improved pharmaceutical protein, a much more elaborate screen may be required. In such cases, molecular breeding of multiple sequences from natural diversity is the best way to generate high-quality libraries that cover a very large area of mostly functional sequence space, so that very few variants need to be tested to obtain the required changes. For example, in the α-interferon example, only sixty-eight assays were used to obtain a significant functional improvement in the first cycle of shuffling. The interferon and subtilisin examples previously described show that the libraries created by molecular breeding can be of unusually high quality. This general approach makes it feasible to perform a small number of assays directly in complex conditions that correlate closely with the final commercial application. We foresee screening small, high-quality libraries of clones directly in whole transgenic plants or animals, especially for whole organism traits for which assays cannot otherwise be obtained, such as yield, drought resistance, or disease resistance.

V. Beyond Proteins

A. *Molecular Breeding of Multigene Phenotypes*

Although most of the previously mentioned methods and examples have focused on improvement of the isolated protein product of a single gene, whole cell biocatalysts make up the majority of industrial biocatalysts. Industrial microorganisms effect the multistep conversion of renewable feedstocks to high value chemical products in a single reactor and comprise a multibillion dollar industry. Fermentation products range from commodity chemicals such as ethanol, organic acids, and amino acids, to high-value small molecule pharmaceuticals, protein pharmaceuticals, and industrial enzymes. Similar to enzymes, whole cell biocatalysts isolated from nature seldom demonstrate the required properties to function under the constraints of a commercial process and thus require specific improvements, such as increased yield of desired products, removal of unwanted co-metabolites, improved utilization of inexpensive carbon and nitrogen sources, and adaptation to fermenter conditions. Success in bringing biocatalytic processes to market and competing in those markets relies on the ability to continuously improve the biocatalyst. The scientific and commercial efforts to understand and manipulate specific functions of whole cells have created the disciplines of metabolic engineering and industrial strain improvement.

Current strategies for strain improvement rely on the empirical and iterative modification of fermenter conditions and genetic manipulation of the whole cell biocatalyst. The genetic manipulation of industrial microorganisms has traditionally taken two paths: the rational approach of metabolic engineering and the empirical approach of classic strain improvement. Although years of intensive research have yielded the genetic tools and information database required to attempt the calculated manipulation of a number of established industrial organisms, metabolic engineering suffers from its reliance on the assumptions of a rational approach. Furthermore, the method and experience gained is typically species-specific and not easily transferred to newly discovered or poorly characterized microorganisms. For these reasons, the most widely practiced strategy is classic strain improvement, which employs random point mutagenesis (chemical or UV) of the producing strain and screening for mutants that have improved properties. Classic strain improvement is robust but suffers from limitations that are typical of iterated point mutagenesis (described above). Molecular breeding, which simulates classical breeding, accelerates whole cell improvement by removing the limitations of metabolic engineering and classical strain improvement.

The strategies of metabolic engineering generally fall into three classes: (1) enhancing the flux through a desired metabolic pathway by amplifying the expression of genes encoding ''rate limiting'' enzymes and those resistant to feedback inhibition (see Jetten *et al.,* 1994); (2) the introduction of exogenous genes, which convert a metabolite of the host organism to a desirable chemical at a viable yield (see Cameron and Tong, 1993); and (3) decreasing the diversion of chemical precursors by the disruption of genes encoding competing pathways. All these approaches closely resemble structure-based design of proteins in that they rely on a great deal of information and are often limited by invalid assumptions. The interpretation of biological data is dominated by information considered ''known'' about the system under investigation. This occurs even though the ''known'' data set is intrinsically incomplete due to the complexity of biological systems. Recent metabolic engineering studies have demonstrated that cellular physiology is extremely robust, and that well-conceived genetic perturbations often result in little or no change in phenotype. Even severe changes to primary metabolism, such as the deletion of the genes encoding pyruvate kinase, have been shown to have negligible effects on primary metabolic fluxes or growth rate in *E. coli* (Sauer *et al.,* 1999). Complex biological systems from single enzymes to whole cells continue to resist rational manipulation but succumb to empirical approaches, such as mutagenesis and screening, which rely on few assumptions.

Each of the three strategies of metabolic engineering has demonstrated value, yet each is limited by its assumptions and the ''cut and paste'' nature of genetic engineering. Overexpressing gene(s) believed to represent a rate limiting step or eliminating feedback regulation of pathway enzyme(s) can be a productive means of enhancing flux through desired pathways (class 1 above). However, this approach often results in only a small increase in rate since other genes affecting the pathway become rate limiting. The term ''rate-limiting step'' is misleading since the rate through a metabolic pathway is generally limited by a collection of enzymes rather than a single enzyme step (Fell, 1998). Metabolic networks are tightly controlled and have evolved to prevent the unnecessary buildup of toxic or useless intermediates. Participating enzymes function at similar rates and under similar conditions to avoid these scenarios, and more than a single enzyme in a given pathway may be under feedback regulation. For example, the biosynthetic enzymes of the aspartate derived amino-acid pathway are under multiple levels of regulation (Eikmanns *et al.,* 1993). Therefore, a small increase in metabolic flux resulting from gene overexpression may be accompanied by a buildup of undesired or detrimental intermediates. Molecular breeding of the genes that encode the pathway enzymes followed by screening the resulting libraries for the desired phenotype provides a direct route to unbiased genetic solutions. This approach allows the improvement of the individual components of the system—for example, improving expression balance within the pathway, eliminating feedback inhibition, improving k_{cat} and K_m for the pathway enzymes, and adaptation to the cellular conditions imparted by the bioprocess. This strategy assumes only that a genetic solution exists within the DNA that is shuffled.

The cloning of heterologous genes to generate new metabolic pathways is one of the most powerful methods for generating new biocatalysts (class 2 above), but poor functioning of the cloned genes often hampers the success of this approach. Genes and gene products have adapted to function in the environment of their native hosts, and these environments are specific to the organisms and their ecological niches. For example, enzymes from thermophilic organisms do not function well in mesophilic hosts. Heterologous genes may be poorly expressed and the encoded polypeptides may not fold properly. Basic genetic elements and the identity of the primary metabolites may be similar between different organisms, but the physical and chemical states of the cells can be significantly different. The concentration of metabolites, pH, temperature, and ionic strength will differ, each influencing the optimal performance of an enzyme; further, the structure of macromolecules with which an enzyme might interact will differ, compromising functional interactions. Thus, a metabolic pathway transplanted from one organism to another may not function optimally. Indeed, the cytoplasmic state of a cell under the conditions of fermentation will be different from that experienced in its natural environment, and even a native pathway may not function optimally. Shuffling heterlogous or native genes and screening them for performance under the desired bioprocess conditions provides a means to identify variants of those genes that have adapted to the new cellular environment and are functioning optimally. The ability of DNA shuffling to alter the substrate preference of enzymes also allows one to access promiscuous activities of enzymes and evolve them to function productively in the context of new metabolic pathways.

The deletion of competing pathways is also a productive route to increasing flux through a desired pathway or at least eliminating potential contaminating products (class 3 above) (Hols *et al.,* 1999). However, the removal of a known pathway may be insufficient to divert flux through the desired pathway, since flux may be limited by either the kinetic parameters of the pathway enzymes or by external factors. Further, the competing pathway may be essential and its elimination may not be an option. The goal is to divert maximal flux down the desired pathway while maintaining only the necessary flux through any competing pathway. Simultaneous shuffling of both pathways should produce an optimal balance in which flux through the desired pathway is maximized, while maintaining the minimal necessary flux through the competing pathway(s) to allow survival. An intrinsic value of the directed evolution approach is that it allows one to find this balance within a complex system. This often is not possible in a straight metabolic engineering ''all-or-nothing'' strategy.

DNA shuffling has been demonstrated to improve the heterologous expression of proteins, alter the substrate specificity of enzymes, and improve the function and stability of enzymes under a variety of extreme environmental conditions. Improvement of single genes by DNA shuffling results in the alteration of the expression, structure, and function of the gene product. In contrast, improvement of metabolic pathways by DNA shuffling results in the alteration of the individual genes (as above) as well as complex interactions of the gene products with each other and the cellular environment. In this way, DNA shuffling complements the strategies of metabolic engineering and provides access to the complex genetic solutions required of strain improvement goals. Crameri *et al.* demonstrated the productivity of this approach by the evolution of the *Staphylococcus aureus* arsenate resistance operon to impart increased resistance to arsenate in *E. coli* (Crameri *et al.* 1997). The pathway consisted of three genes encoding an arsenate reductase, an arsenite efflux pump, and a regulatory protein. Previous rational work suggested that any improvements required would be found in the arsenate reductase. After three rounds of shuffling and screening, a variant of the operon imparting a resistance to 0.5 M arsenate was identified (a forty-fold improvement). Analysis of the new operon identified two major surprises: most of the thirteen mutations were clustered in the efflux pump with no mutations found within the coding region of the reductase, and the originally episomal plasmid had integrated into the chromosome and this shuffling-dependent integration was shown to contribute a large part of the improvement. These data emphasize the complex and non-intuitive solutions that arise from a directed evolution approach and demonstrate the utility of molecular breeding to optimize the function of a complete metabolic pathway.

B. Genome Shuffling

Long before molecular geneticists began tinkering with the structure and function of proteins and metabolic pathways, researchers were manipulating the performance of industrial microorganisms by classic strain improvement. These classical approaches remain an important part of all strain improvement programs, primarily because they are robust and reproducibly yield new strains with slightly improved phenotype. Although metabolic engineering requires a great deal of information and molecular tools, classical strain improvement requires only a starting organism, a mutagen, and a good screen for improvement. Cellular phenotypes are complex and are influenced by many more genes than those that are recognized as ''necessary and sufficient.'' Although improvements in phenotype may be accessible by variations within a defined set of genes, other elements distributed throughout the genome may have equal or greater influence. Again, an analogy with protein design is relevant. The improvement of enzyme function is most productive when the entire structural gene is targeted as opposed to only those regions known to encode the active site. Similarly, the most productive mode of improving whole cell biocatalysts is through the evolution of the cell's entire genome. The robust nature of classical strain improvement lies in the fact that it is unbiased and can address complex, distributed phenotypes. The superior performance of classical strain improvement over metabolic engineering is a testament to this fact. The limitations to classical strain improvement are the same as those of sequential point mutagenesis—the process is asexual. Improvements are small and one can accumulate only one beneficial mutation at a time. Genome shuffling incorporates recombination into the strain improvement process and thereby significantly accelerates the process. It provides the means to recombine the genomic information from many strains so that the useful alleles from all of them can be combined into a single superior organism. Useful mutations are combined and deleterious mutations are replaced with wild-type sequence. Instead of accumulating a single beneficial genetic event per cycle of mutagenesis and screening, evolution occurs via large leaps by creating complex combinations of multiple mutations.

VI. Concluding Remarks

Our preferred method for molecular breeding involves recombination of homologous genes obtained from nature, in order to permutate the proven diversity. These libraries are high quality (rich in functional sequences) because the variations have been prescreened for function in nature and phenotypically diverse. In rare cases when adequate natural diversity is not available, such as when homologous sequences are not known or if the target is a small segment of a protein, the sequence diversity must be generated artificially. Typical methods include random mutagenesis of a single DNA sequence followed by screening for the best mutations, various kinds of synthetic oligonucleotide cassette mutagenesis of a small part of a protein, or mutations that were suggested based on molecular modeling of the protein's structure.

However, regardless of the source of variation, recombination by DNA shuffling is the most effective method for creating higher order combinations of previously selected mutations, whether the targets are single genes, pathways, or whole genomes.

Acknowledgments

We thank Lori Giver, Seran Kim, and Phil Patten for critical input; and Mark Welch for the bubble plot of subtilisins.

References

Arkin, A. P., and Youvan, D. C. (1992). ''Optimizing nucleotide mixtures to encode specific subsets of amino acids for semi-random mutagenesis.'' *Bio/technology,* **10,** 297–300.

Arnold, F. (1998a). ''Design by directed evolution.'' *Acc. Chem. Res.,* **31,** 125–131.

Arnold, F. H. (1998b). ''When blind is better: protein design by evolution.'' *Nature Biotechnology,* **16,** 617–618.

Arnold, G. E., and Ornstein, R. L. (1997). ''Molecular dynamics study of time-correlated protein domain motions and molecular flexibility: cytochrome P450BM-3.'' *Biophys J.,* **73,** 1147–1159.

Babbitt, P. C., Mrachko, G. T., Hasson, M. S., Huisman, G. W., Kolter, R., Ringe, D., Petsko, G. A., Kenyon, G. L., and Gerlt, J. A. (1995). ''A functionally diverse enzyme superfamily that abstracts the alpha protons of carboxylic acids.'' *Science,* **267**(5201), 1159–1161.

Baca, M., Scanlan, T. S., Stephenson, R. C., and Wells, J. A. (1997). ''Phage display of a catalytic antibody to optimize affinity for transition-state analog binding.'' *Proc. Natl. Acad. Sci. USA,* **94**(19), 10063–10068.

Ballinger, M. D., Tom, J., and Wells, J. A. (1996). ''Furilisin: a variant of subtilisin BPN' engineered for cleaving tribasic substrates.'' *Biochemistry,* **35**(42), 13579–13585.

Bhosale, S. H., Rao, M. B., and Deshpande, V. V. (1996). ''Molecular and industrial aspects of glucose isomerase.'' *Microbiol Rev.,* **60**(2), 280–300.

Black, M. E., and Loeb, L. A. (1993). ''Identification of important residues within the putative nucleoside binding site of HSV-1 thymidine kinase by random sequence selection: analysis of selected mutants in vitro.'' *Biochemistry,* **32,** 11618–11626.

Black, M. E., Newcomb, T. G., Wilson, H. M., and Loeb, L. A. (1996). ''Creation of drug-specific herpes simplex virus type 1 thymidine kinase mutants for gene therapy.'' *Proc. Natl. Acad. Sci. USA,* **93,** 3525–3529.

Bornscheuer, U. T., Altenbuchner, J., and Meyer, H. H. (1998). ''Directed evolution of an esterase for the stereoselective resolution of a key intermediate in the synthesis of epithilones.'' *Biotechnology and Bioengineering,* **58,** 554–559.

Bott, R., and Betzel, C. (1996). *Subtilisin Enzymes,* Plenum Press, New York.

Bryan, P. N., Rollence, M. L., Pantoliano, M. W., Wood, J., Finzel, B. C., Gilliland, G. L., Howard, A. J., and Poulos, T. L. (1986). ''Proteases of enhanced stability: characterization of a thermostable variant of subtilisin.'' *Proteins,* **1,** 326–334.

Burbank, L., Whitson, J., John, R., Williams, H. S., and Luther Burbank Society. (1914). *Luther Burbank, his methods and discoveries and their practical application,* Luther Burbank Press, New York; London.

Cadwell, R. C., and Joyce, G. F. (1992). ''Randomization of genes by PCR mutagenesis.'' *PCR Methods Appl.,* **2**(1), 28–33.

Cadwell, R. C., and Joyce, G. F. (1994). ''Mutagenic PCR.'' *PCR Methods Appl.,* **3**(6), S136–S140.

Cameron, D. C., and Tong, I.-T. (1993). ''Cellular and metabolic engineering. An overview.'' *Appl. Biochem. Biotechnol.,* **38,** 105–140.

Chang, C. C., Chen, T. T., Cox, B. W., Dawes, G. N., Stemmer, W. P., Punnonen, J., and Patten, P. A. (1999). ''Evolution of a cytokine using DNA family shuffling.'' *Nat. Biotechnol.,* **17**(8), 793–797.

Chen, G., Dubrawsky, I., Mendez, P., Georgiou, G., and Iverson, B. L. (1999). ''*In vitro* scanning saturation mutagenesis of all the specificity determining residues in an antibody binding site.'' *Protein Eng.,* **12**(4), 349–356.

Chen, K., and Arnold, F. H. (1991). ''Enzyme engineering for nonaqueous solvents: random mutagenesis to enhance activity of subtilisin E in polar organic media.'' *Bio/Technology,* **9,** 1073–1077.

Cherry, J. R., Lamsa, M. H., Schneider, P., Vind, J., Svendsen, A., Jones, A., and Pedersen, A. H. (1999). ''Directed evolution of a fungal peroxidase.'' *1999,* **17,** 379–384.

Christians, F. C., Scapozza, L., Crameri, A., Folkers, G., and Stemmer, W. P. C. (1999). ''Directed evolution of thymidine kinase for AZT phosphorylation using DNA family shuffling.'' *Nature Biotechnol.,* **17.**

Crameri, A., Dawes, G., Rodriguez, E., Silver, S., and Stemmer, W. P. C. (1997). ''Molecular evolution of an arsenate detoxification pathway by DNA shuffling.'' *Nature Biotechnology,* **15,** 436–438.

Crameri, A., Raillard, S.-A., Bermudez, E., and Stemmer, W. P. C. (1998). ''DNA shuffling of a family of genes from diverse species accelerates directed evolution.'' *Nature,* **391,** 288–291.

Crameri, A., Whitehorn, E. A., Tate, E., and Stemmer, W. P. C. (1996). ''Improved green fluorescent protein by molecular evolution using DNA shuffling.'' *Nature Biotechnology,* **14,** 315–319.

Cunningham, B. C., and Wells, J. A. (1987). ''Improvement in the alkaline stability of subtilisin using an efficient random mutagenesis and screening procedure.'' *Protein Eng.,* **1**(4), 319–325.

Dahiyat, B. I., and Mayo, S. L. (1997). ''De novo protein design: fully automated sequence selection.'' *Science,* **278,** 82–87.

Darwin, C. (1859). *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life,* John Murray, London.

Daugherty, P. S., Chen, G., Olsen, M. J., Iverson, B. L., and Georgiou, G. (1998). ''Antibody affinity maturation using bacterial surface display.'' *Protein Eng.,* **11**(9), 825–32.

Daugherty, P. S., Olsen, M. J., Iverson, B. L., and Georgiou, G. (1999). ''Development of an optimized expression system for the screening of antibody libraries displayed on the *Escherichia coli* surface.'' *Protein Eng.,* **12**(7), 613–621.

Delagrave, S., Goldman, E. R., and Youvan, D. C. (1993). ''Recursive ensemble mutagenesis.'' *Protein Eng.,* **6**(3), 327–371.

Delagrave, S., and Youvan, D. C. (1993). ‘‘Searching sequence space to engineer proteins: exponential ensemble mutagenesis.’’ *Biotechnology (NY)*, **11**(13), 1548–1552.

Diamond, J. M. (1997). *Guns, germs, and steel: the fates of human societies,* W. W. Norton, New York.

Eikmanns, B. J., Eggeling, L., and Sahm, H. (1993). ‘‘Molecular aspects of lysine, threonine, and isoleucine biosynthesis in *Corynebacterium glutamicum.*’’ *Antonie Van Leeuwenhoek,* **64**(2), 145–163.

Fell, D. A. (1998). ‘‘Increasing the flux in metabolic pathways: A metabolic control analysis perspective.’’ *Biotechnol Bioeng,* **58**(2-3), 121–124.

Forrest, S. (1993). ‘‘Genetic algorithms: principles of natural selection applied to computation.’’ *Science,* **261,** 872–878.

Georgiou, G., Stathopoulos, C., Daugherty, P. S., Nayak, A. R., Iverson, B. L., and Curtiss, R., 3rd. (1997). ‘‘Display of heterologous proteins on the surface of microorganisms: from the screening of combinatorial libraries to live recombinant vaccines.’’ *Nat. Biotechnol.,* **15**(1), 29–34.

Gibson, J. M. (1989). ‘‘Simulated evolution and artificial selection.’’ *BioSystems,* **23,** 219–229.

Giver, L., Gershenson, A., Freskgard, P. O., and Arnold, F. H. (1998). ‘‘Directed evolution of a thermostable esterase.’’ *Proc. Natl. Acad. Sci. USA,* **95,** 12809–12813.

Graycar, T., Knapp, M., Ganshaw, G., Dauberman, J., and Bott, R. (1999). ‘‘Engineered *Bacillus lentus* subtilisins having altered flexibility.’’ *J. Mol. Biol.,* **292**(1), 97–109.

Graycar, T. P., Bott, R. R., Caldwell, R. M., Dauberman, J. L., Lad, P. J., Power, S. D., Sagar, I. H., Silva, R. A., Weiss, G. L., Woodhous, L. R., and Estell, D. A. (1992). ‘‘Altering the proteolytic activity of stubtilisin through protein engineering.’’ Enzyme Engineering XI, D. S. Clark and D. A. Estell, eds., The New York Academy of Sciences, New York, 71–79.

Hagedorn, S., and Kaphammer, B. (1994). ‘‘Microbial biocatalysis in the generation of flavor and fragrance chemicals.’’ *Ann. Rev. Microbiol,* **48,** 773–800.

Hanes, J., Jermutus, L., Schaffitzel, C., and Pluckthun, A. (1999). ‘‘Comparison of *Escherichia coli* and rabbit reticulocyte ribosome display systems.’’ *FEBS Lett,* **450**(1–2), 105–110.

Hanes, J., Jermutus, L., Weber-Bornhauser, S., Bosshard, H. R., and Pluckthun, A. (1998). ‘‘Ribosome display efficiently selects and evolves high-affinity antibodies *in vitro* from immune libraries.’’ *Proc. Natl. Acad. Sci. USA,* **95**(24), 14130–14135.

Hanes, J., and Pluckthun, A. (1997). ‘‘*In vitro* selection and evolution of functional proteins by using ribosome display.’’ *Proc. Natl. Acad. Sci. USA,* **94**(10), 4937–4942.

Hastrup, S., Branner, S., Norris, F., Petersen, S. B., Norskov-Lauridsen, L., Jensen, V. J., and Aaslyng, D. (1989). ‘‘Mutated subtilisin genes.’’ PCT Patent Appl. WO 8906279, Novo Industries, Denmark.

Hodits, R. A., Nimpf, J., Pfistermueller, D. M., Hiesberger, T., Schneider, W. J., Vaughan, T. J., Johnson, K. S., Haumer, M., Kuechler, E., Winter, G., *et al.* (1995). ‘‘An antibody fragment from a phage display library competes for ligand binding to the low density lipoprotein receptor family and inhibits rhinovirus infection.’’ *J. Biol. Chem.,* **270**(41), 24078–24085.

Holland, J. H. (1975). *Adaptation in Natural and Artificial Systems,* Univ. of Michigan Press, Ann Arbor, MI.

Hols, P., Kleerebezem, M., Schanck, A. N., Ferain, T., Hugenholtz, J., Delcour, J., and de Vos, W. M. (1999). ‘‘Conversion of *Lactococcus lactis* from homolactic to homoalanine fermentation through metabolic engineering.’’ *Nat. Biotechnol.,* **17**(6), 588–592.

Janda, K. D., Lo, C. H., Li, T., Barbas, C. F., 3rd, Wirsching, P., and Lerner, R. A. (1994). ''Direct selection for a catalytic mechanism from combinatorial antibody libraries.'' *Proc. Natl. Acad. Sci. USA,* **91**(7), 2532–2536.

Janes, L. E., and Kazlauskas, R. J. (1997). ''Quick E. A fast spectrophotometric method to measure the enatntioselectivity of hydrolases.'' *J. Org. Chem.,* **62,** 4560–4561.

Jespers, L., Jenne, S., Lasters, I., and Collen, d. (1997). ''Epitope mapping by negative selection of randomized antigen libraries displayed on filamentous phage.'' *J. Mol. Biol.,* **269,** 704–718.

Jetten, M. S., Follettie, M. T., and Sinskey, A. J. (1994). ''Metabolic engineering of *Corynebacterium glutamicum.*'' *Ann. NY Acad. Sci.,* **721,** 12–29.

Joo, H., Lin, Z., and Arnold, F. H. (1999). ''Laboratory evolution of peroxide-mediated cytochrome P450 hydroxylation [see comments].'' *Nature,* **399**(6737), 670–673.

Kano, H., Taguchi, S., and Momose, H. (1997). ''Cold adaptation of a mesophilic serine protease, subtilisin, by *in vitro* random mutagenesis.'' *Appl. Microbiol. Biotechnol.,* **47,** 46–51.

Kelly, K. (1994). *Out of Control: the rise of neo-biological civilization,* Addison-Wesley, Menlo Park, California.

Kobayashi, M., Nagasawa, T., and Yamada, H. (1992). ''Enzymatic synthesis of acrylamide: a success story not yet over.'' *Trends Biotechnol.,* **10**(11), 402–408.

Kunkel, T. A., Bebenek, K., and McClary, J. (1991). ''Efficient site-directed mutagenesis using uracil-containing DNA.'' *Methods Enzymol.,* **204,** 125–139.

Lobkovsky, E., Moews, P. C., Liu, H., Zhao, H., Frere, J. M., and Knox, J. R. (1993). ''Evolution of an enzyme activity: crystallographic structure at 2-A resolution of cephalosporinase from the *ampC* gene of *Enterobacter cloacae* P99 and comparison with a class A penicillinase.'' *Proc. Natl. Acad. Sci. USA,* 90(23), 11257–11261.

Low, N. M., Holliger, P. H., and Winter, G. (1996). ''Mimicking somatic hypermutation: affinity maturation of antibodies displayed on bacteriophage using a bacterial mutator strain.'' *J. Mol. Biol.,* **260,** 359–368.

Manivasakam, P., Weber, S. C., McElver, J., and Schiestl, R. H. (1995). ''Micro-homology mediated PCR targeting in *Saccharomyces cerevisiae.*'' *Nucleic Acids Res.,* **23**(14), 2799–2800.

Matic, I., Rayssiguier, C., and Radman, M. (1995). ''Interspecies gene exchange in bacteria: the role of SOS and mismatch repair systems in evolution of species.'' *Cell,* **80,** 507–515.

Matsumura, I., Wallingford, J. B., Surana, n. K., Vize, P. D., and Ellington, A. D. (1999). ''Directed evolution of the surface chemistry of the reporter enzyme β-glucuronidase.'' *Nature Biotechnology,* **17,** 696–701.

Matsumura, M., Yasumura, S., and Aiba, S. (1986). ''Cumulative effect of intragenic amino-acid replacements on the thermostability of a protein.'' *Nature,* **323,** 356–358.

Mattheakis, L. C., Bhatt, R. R., and Dower, W. J. (1994). ''An *in vitro* polysome display system for identifying ligands from very large peptide libraries.'' *Proc. Natl. Acad. Sci. USA,* **91**(19), 9022–9026.

Minshull, J., and Stemmer, P. (1999). ''Protein evolution by molecular breeding.'' *Curr. Opin. Chem. Biol.,* **3**(3), 284–290.

Moore, J. C., Jin, H. M., Kuchner, O., and Arnold, F. H. (1997). ''Strategies for the *in vitro* evolution of protein function: enzyme evolution by random recombination of improved sequences.'' *J. Mol. Biol.,* **272**(3), 336–347.

Muller, H. J. (1964). ''The relation of recombination to mutational advance.'' *Mutation Research,* **1,** 2–9.

Naki, D., Paech, C., Granshaw, G., and Schellenberger, V. (1998). ''Selection of a subtilisin-hyperproducing *Bacillus* in a highly structured environment.'' *Appl. Microbiol. Biotechnol.,* **49,** 290–294.

Narinx, E., Baise, E., and Gerday, C. (1997). ''Subtilisin from psychrophilic antarctic bacteria: characterization and site-directed mutagenesis of residues possibly involved in the adaptation to cold.'' *Protein Engineering,* **10,** 1271–1279.

Nedwin, G. (1997). ''Using enzymes as benign substitutes for synthetic chemicals and harch conditions in industrial processes.'' Biotechnology in the Sustainable Environment, G. Sayler, J. Sanseverino, and K. Davis, eds., Plenum Press, New York, 13–32.

Ness, J. E., Welch, M., Giver, L., Bueno, M., Cherry, J. R., Borchert, T. V., Stemmer, W. P., and Minshull, J. (1999). ''DNA shuffling of subgenomic sequences of subtilisin.'' *Nat. Biotechnol.,* **17**(9), 893–896.

Ogunseitan, O. A. (1995). ''Bacterial genetic exchange in nature.'' *Science Progress,* **78**(3), 183–204.

Palzkill, T., and Botstein, D. (1992). ''Identification of amino acid substitutions that alter the substrate specificity of TEM-1 beta-lactamase.'' *J. Bacteriol.,* **174,** 5237–5243.

Patkar, S., Vind, J., Kelstrup, E., Christensen, M. W., Svendsen, A., Borch, K., and Kirk, O. (1998). ''Effect of mutations in *Candida antarctica* B lipase.'' *Chem. Phys. Lipids,* **93,** 95–101.

Patten, P. A., Howard, R. J., and Stemmer, W. P. C. (1997). ''Applications of DNA shuffling to pharmaceuticals and vaccines.'' *Curr. Opin. Biotechnol.,* **8,** 724–733.

Reidhaar-Olson, J. F., Bowie, J. U., Breyer, R. M., Hu, J. C., Knight, K. L., Lim, W. A., Mossing, M. C., Parsell, D. A., Shoemaker, K. R., and Sauer, R. T. (1991). ''Random mutagenesis of protein sequences using oligonucleotide cassettes.'' *Methods Enzymol.,* **208,** 564–586.

Roberts, R. W., and Szostak, J. W. (1997). ''RNA-peptide fusions for the *in vitro* selection of peptides and proteins.'' *Proc. Natl. Acad. Sci. USA.* **94**(23), 12297–12302.

Roeder, G. S. (1997). ''Meiotic chromosomes: it takes two to tango.'' *Genes Dev.,* **11**(20), 2600–2621.

Rubingh, D. N. (1997). ''Protein engineering from a bioindustrial point of view.'' *Curr. Opin. Biotechnol.,* **8,** 417–422.

Russel, A. J., and Fersht, A. R. (1987). ''Rational modification of enzyme catalysis by engineering surface charge.'' *Nature,* **328,** 496–500.

Sauer, U., Lasko, D. R., Fiaux, J., Hochuli, M., Glaser, R., Szyperski, T., Wuthrich, K., and Bailey, J. E. (1999). ''Metabolic flux ratio analysis of genetic and environmental modulations of *Escherichia coli* central carbon metabolism.'' *J. Bacteriol.,* **181**(21), 6679–6688.

Scanlan, T. S., and Reid, R. C. (1995). ''Evolution in action.'' *Chem. Biol.,* **2,** 71–75.

Shafikhani, S., Siegel, R. A., Ferrari, E., and Schellenberger, V. (1997). ''Generation of large libraries of random mutants in Bacillus subtilis by PCR-based plasmid multimerization.'' *Biotechniques,* **23**(2), 304–310.

Shao, Z., and Arnold, F. H. (1996). ''Engineering new functions and altering existing functions.'' *Curr. Op. Structural Biol.,* 6, 513–518.

Shao, Z., Zhao, H., Giver, L., and Arnold, F. H. (1998). ''Random-priming *in vitro* recombination: an effective tool for directed evolution.'' *Nucl. Acids Res.,* **26,** 681–683.

Shaw, A., Bott, R., and Day, A. G. (1999). ''Protein engineering of α-amylase for low pH performance.'' *Curr. Opin. Biotechnol.,* **10**(4), 349–352.

Shoichet, B. K., Baase, W. A., Kuroki, R., and Matthews, B. W. (1995). ''A relationship between protein stability and protein function.'' *Proc. Natl. Acad. Sci. USA,* **92,** 452–456.

Smiley, J. A., and Benkovic, S. J. (1994). ''Selection of catalytic antibodies for a biosynthetic reaction from a combinatorial cDNA library by complementation of an auxotrophic *Escherichia coli:* antibodies for orotate decarboxylation.'' *Proc, Natl. Acad. Sci. USA,* **91**(18), 8319–8323.

Smith, G. P., Patel, S. U., Windass, J. D., Thornton, J. M., Winter, G., and Griffiths, A. D. (1998). ''Small binding proteins selected from a combinatorial repertoire of knottins displayed on phage.'' *J. Mol. Biol.,* **277**(2), 317–332.

Smith, J. M., Dowson, C. G., and Spratt, B. G. (1991). ''Localized sex in bacteria.'' *Nature,* **349,** 29–31.

Soong, N.-W., Nomura, L., Pekrun, K., Reed, M., Sheppard, L., Dawes, G., and Stemmer, W. P. C. (2000). ''Molecular Breeding of Viruses.'' *Nature Genetics.* In press.

Spiller, B., Gershenson, A., Arnold, F. H., and Stevens, R. C. (1999). ''A structural view of evolutionary divergence.'' *Proc. Natl. Acad. Sci. USA,* **96**(22), 12305–12310.

Stemmer, W. P. (1994a). ''DNA shuffling by random fragmentation and reassembly: *in vitro* recombination for molecular evolution.'' *Proc. Natl. Acad. Sci. USA,* **91,** 10747–10751.

Stemmer, W. P. (1994b). ''Rapid evolution of a protein *in vitro* by DNA shuffling.'' *Nature,* **370,** 389–391.

Suzuki, M., Christians, F. C., Kim, B., Skandalis, A., Black, M. E., and Loeb, L. A. (1996). ''Tolerance of different proteins for amino acid diversity.'' *Mol. Divers,* **2**(1–2), 111–118.

Taguchi, S., Ozaki, A., and Momose, H. (1998). ''Engineering of a cold-adapted protease by sequential random mutagenesis and a screening system.'' *Appl. Environmental Microbiol.,* **64**(2), 492–495.

van der Meer, J. R., de Vos, W. M., Harayama, S., and Zehnder, A. J. (1992). ''Molecular mechanisms of genetic adaptation to xenobiotic compounds.'' *Microbiol Rev.,* **56**(4), 677–694.

Wells, J. A. (1990). ''Additivity of mutational effects in proteins.'' *Biochemistry,* **29**(37), 8509–8517.

Wells, J. A., and Estell, D. A. (1988). ''Subtilisin—an enzyme designed to be engineered.'' *TIBS,* **13,** 291–297.

Winter, G., Griffiths, A. D., Hawkins, R. E., and Hoogenboom, H. R. (1994). ''Making antibodies by phage display technology.'' *Ann. Rev. Immunol,* **12,** 433–455.

Wong, S. L., Ye, R., and Nathoo, S. (1994). ''Engineering and production of streptokinase in a *Bacillus subtilis* expression-secretion system.'' *Appl. Environ. Microbiol.,* **60**(2), 517–523.

Wu, X. C., Ye, R., Duan, Y., and Wong, S. L. (1998). ''Engineering of plasmin-resistant forms of streptokinase and their production in *Bacillus subtilis*: streptokinase with longer functional half-life.'' *Appl. Environ. Microbiol.,* **64**(3), 824–829.

Yang, M. M. (1994). ''Digital imaging spectroscopy of microbial colonies.'' *Am. Biotechnol Lab,* **12**(6), 18–20.

You, L., and Arnold, F. H. (1996). ''Directed evolution of subtilisin E in *Bacillus subtilis* to enhance total activity in aqueous dimethylformamide.'' *Protein Eng.,* **9,** 77–83.

Zhang, J.-H., Dawes, G., and Stemmer, W. P. C. (1997). ''Directed evolution of a fucosidase from a galactosidase by DNA shuffling and screening.'' *Proc. Natl. Acad. Sci. USA,* **94,** 4505–4509.

Zhao, H., and Arnold, F. (1997a). ''Combinatorial protein design: strategies for screening protein libraries.'' *Current Opinion in Structural Biology,* **7,** 480–485.

Zhao, H., and Arnold, F. H. (1997b). ''Functional and nonfunctional mutations distinguished by random recombination of homologous genes.'' *Proc. Natl. Acad. Sci. USA,* **94,** 7997–8000.

Zhao, H., and Arnold, F. H. (1997c). ''Optimization of DNA shuffling for high fidelity recombination.'' *Nucleic Acids Res.,* **25**(6), 1307–1308.

Zhao, H., Giver, L., Shao, Z., Affholter, J. A., and Arnold, F. H. (1998). ''Molecular evolution by staggered extension process (StEP) *in vitro* recombination.'' *Nature Biotechnol,* **16,** 258–261.

ANALYSIS OF LARGE LIBRARIES OF PROTEIN MUTANTS USING FLOW CYTOMETRY

By GEORGE GEORGIOU

Department of Chemical Engineering and Institute for Cell and Molecular Biology,
University of Texas, Austin, Texas 78712

I. Introduction

One of the most intriguing problems in directed protein evolution is determining the optimal strategy for exploring the sequence space and isolating gain-of-function, change-of-function, or stability mutants. The most widely adopted experimental strategy is the iterative search of libraries containing low rates of random nucleotide substitutions (Arnold, 1998). This approach has been dictated by both experimental limitations as well as theoretical considerations: Assaying for enzyme function is generally tedious and represents the rate limiting step in library screening. Until recently, the number of independent clones that could be assayed for enzymatic function by liquid or plate assays was limited to around 5×10^5 clones (Joo *et al.,* 1999). If a low rate of mutagenesis is used, a large fraction of all the possible amino-acid substitutions (''sequence space'') may be represented in a library that is still small enough to be screened by conventional agar plate or microtiter well assays. In addition, a low rate of mutagenesis is considered necessary to maintain the fraction of deleterious mutations at a tolerable level (Kuchner and Arnold, 1998).

The iterative screening of relatively small libraries of mutants with a low frequency of nucleotide substitutions has proved to be extremely effective for the functional improvement of numerous proteins and has literally changed the way we think about protein design (Arnold, 2000). However, as high throughput screening methodologies are becoming

0065-3233/00 $35.00

increasingly more powerful, the analysis and screening of libraries several orders of magnitude larger than was possible only two or three years ago are now within reach. Larger swaths of the protein sequence space can be accessed experimentally and the frequency of mutants exhibiting different phenotypes can be determined quantitatively and in a systematic fashion. The analysis of larger and larger libraries may well push directed evolution to frontiers that one can barely imagine today.

A number of high throughput screening techniques can be used for protein engineering purposes. Among them, flow cytometry stands out as an extremely useful tool for the quantitative analysis of protein libraries and for the isolation and characterization of very rare clones (Shusta *et al.,* 1999a; Daugherty *et al.,* 2000a). High-end flow cytometers (such as Cytomation's MoFlo™) can sort up to about 3 × 10^8 cells per hour and, thus, the analysis of billions of clones can be realized within a short time. Unlike techniques such as phage display, flow cytometry provides information on the activity of each and every clone in a library. As a result, the distribution of function within the entire population becomes directly available. Clones with very low, intermediate, and high activities can be detected and isolated individually and/or as different pools.

With flow cytometry and other high throughput screening techniques, ultimately, the exploration of the protein sequence space will be limited not by the screening step but rather by how large a library can be constructed. At present, the practical limit for the construction of expression libraries in microorganisms stems from limitations in bacterial transformation and is about 3 × 10^9. Barring some unforeseen breakthrough in bacterial transformation technology (or a herculean effort) this limit is not likely to be surpassed in the near future. *In vitro* protein synthesis may ultimately increase the maximum attainable library size to 10^{14}; however, at present that technology is limited to screening for ligand binding. Will the ability to screen libraries of 10^9–10^{14} clones lead to entirely different strategies for protein evolution? How does the fraction of functional clones in a library vary as a function of the mutation rate? Are there certain protein folds or classes of proteins that are exceptionally tolerant to mutations? The application of the new high throughput screening methodologies to protein libraries is beginning to provide experimental data that will shed light on these issues.

II. Library Screening Technologies

A wide variety of library screening technologies have been developed. Two and three hybrid systems and analogous technologies have proved extremely useful for detecting pairs of interacting proteins or peptides

(Drees, 1999; Pelletier *et al.,* 1999). Very recently, Benkovic and coworkers developed a three-hybrid system that can be used for the isolation of novel protein catalysts (Firestine, *et al.,* 2000). Nonetheless, the application of these techniques to directed protein evolution has so far been rather limited. For protein engineering, the screening of large libraries (10^6 or more clones) can be realized by (a) biological selections; (b) robotic, high throughput colony assays using high density microtiter well plates (>1056 wells/plate); (c) phage display; (d) *in vitro* screening formats utilizing the direct coupling of the RNA template to the newly synthesized polypeptide; and (e) flow cytometry and other techniques that rely on the interrogation of single cells based on fluorescence.

Where applicable, biological selections are a very powerful means for the isolation of gain-of-function or stability mutants. For example, selections have been used to determine which residues in proteins are essential for function or, more recently, for remodeling protein structure (Bowie and Sauer, 1989; Markiewicz *et al.,* 1994; Martinez *et al.,* 1996; Huang *et al.,* 1996; MacBeath *et al.,* 1998). However, an important limitation of biological selections is that only clones expressing proteins that confer a desired phenotype are scored. Thus, no information can be gleaned regarding mutants that are weakly active and cannot confer the phenotype. Moreover, selections are frequently complicated by biological constraints—for example, when a phenotype results not from the activity of a mutant protein *per se* but rather an increase in the expression level or the ability of the mutant to induce a bypass pathway or a stress response.

Miniaturized assays utilize high density microtiter well plates. Arrays of up to 6144 wells with volumes of 1–2 μl or less have been reported (Stockwell *et al.,* 1999). Higher density arrays utilizing nanoliter size samples are being used industrially, but as of this time details have not yet been reported in the scientific literature (Fernandez, 1999). Screening with such small volumes requires robust assays and sophisticated robotic automation for the precise handling of very small volumes, but practically any assay for binding or catalytic activity can be configured for high density screening format. Extremely sensitive assays such as fluorescence cross-correlation spectroscopy, which is useful for the detection of protein activity in femtoliter volumes, have been developed (Koltermann *et al.,* 1998; Winkler *et al.,* 1999). However, the equipment cost and level of technical sophistication required for such screening programs are beyond the reach of most academic laboratories. In industry, miniaturized assays have been used primarily in drug discovery and at present no reports have been published on protein engineering applications.

The screening of large libraries is greatly simplified by establishing a direct physical link between a gene, the protein it encodes, and a desired function. Such a link can be established using a variety of *in vivo* or *in vitro* display technologies (Rodi and Makowski, 1999; Shusta *et al.,* 1999; Roberts, 1999). For *in vivo* methods, the protein is encoded by a recombinant gene and displayed on the surface of a biological particle such as a virus or a whole cell. Currently, the most widely used technique for protein library screening is based on display on the surface of filamentous bacteriophages. In phage display, a gene of interest is fused in-frame to phage genes encoding surface-exposed proteins, most commonly pIII. The gene fusions are translated into chimeric proteins in which the two domains fold independently. Phage displaying a protein with binding affinity for a ligand can be readily enriched by selective adsorption onto immobilized ligand, a process known as ''panning.'' The bound phage is desorbed from the surface, usually by acid elution, and amplified through infection of *E. coli* cells. Usually, four to six rounds of panning and amplification are sufficient to select for phage displaying specific polypeptides, even from very large libraries with diversities up to 10^{11}.[1] Several variations of phage display for the rapid enrichment of clones displaying tightly binding polypeptides have been developed. They include Selectively Infective Phage (SIP) (Spada *et al.,* 1997), where an infective phage particle is generated only if it binds to the ligand and Delayed Infectivity Panning (DIP), a method for screening libraries for binding to polypeptides expressed on the surface of bacterial cells via fusion to an Lpp-OmpA chimera (I. Benhar, private communication).

With phage display, enrichment of the desired clones can only be accomplished through a binding event. Clever strategies are required to link a phage expressing a certain polypeptide to the catalytic turnover of a substrate. Enzymes have been successfully isolated from libraries by selecting for variants that bind transition state analogs or by covalent trapping of catalytically active proteins with suicide inhibitors (e.g., Forrer *et al.,* 1999). Alternatively, the substrate can be covalently linked to the phage via a fusion to pIII, placing it in close proximity to the displayed polypeptide (which is part of the same chimera). Catalytic turnover of the substrate is exploited to either directly ligate the phage onto a suitable solid phase or to generate a product that can be recognized specifically by an affinity matrix, thus resulting in selective enrichment

[1] Libraries of such size have so far only been generated for antibody repertoires. These are constructed by *in vivo* recombination of smaller, primary sublibraries generated by PCR (e.g., Sblattero and Bradbury, 2000). This level of sequence diversity may be attainable for other proteins, as long as they consist of multiple independent folding domains that can tolerate the insertion of recombination sites within the linker region.

(Pedersen *et al.,* 1998; Atwell and Wells, 1999; Demartis *et al.,* 1999). These techniques are useful for enriching catalysts from a large excess of noncatalytic clones but cannot be used to discriminate between proteins exhibiting small or moderate differences in enzymatic activity. Also, phage panning, and for that matter all methods relying on selection via sequential adsorption and desorption, provide information only for the clones that bind a ligand. Therefore, they cannot be used for the quantitative analysis of an entire library. At the end of a phage panning experiment there is no data on the percentage of clones exhibiting a certain level of activity.

In vitro protein synthesis is an excellent means for the facile analysis of large numbers of site-specific protein mutants (Burks *et al.,* 1997; Chen *et al.,* 1999). In addition, several exciting *in vitro* library screening techniques have been developed that completely circumvent the need to use a host cell for protein synthesis. *In vitro* screening technologies employ RNA polymerase for mRNA synthesis from a DNA template. The mRNA is then translated into protein using ribosomal extracts, amino acids, and various cofactors. The advantage of *in vitro* library screening methods is that the library size that can be screened for binding is vast, potentially exceeding 10^{14} different sequences. In theory at least, screening such highly diverse libraries can result in the discovery of entirely new protein folds (Roberts, 1999). Ribosome display, originally described by Mattheakis *et al.* (Mattheakis *et al.,* 1994), and subsequently adopted to the screening of antibody libraries (Hanes *et al.,* 1998; He and Taussig, 1997), relies on the formation of a ternary complex among ribosomes, mRNA, and the polypeptide. Complexes containing folded proteins with a desired specificity are enriched by panning against immobilized ligand. The RNA of the ribosome-mRNA-protein complexes is reverse transcribed to produce the DNA encoding the corresponding protein. Alternatively, a covalent link can be established directly between a nucleic acid sequence and the protein it encodes (Roberts and Szostak, 1997). Covalent RNA-protein complexes are formed via the reactive amino acid analogue puromycin. One important advantage of this technology is that covalent RNA-protein complexes are more stable and can be subjected to harsh screening conditions to select for proteins with increased stability.

Adapting techniques based on *in vitro* protein synthesis to the isolation of enzymes requires establishing a link between a nucleic acid-protein complex and product formation. Methods based on binding, analogous to those developed for phage displayed libraries, may be used to enrich catalysts from noncatalysts. In addition, Tawfik and Griffiths (1998) exploited the aqueous core of reverse micelles as artificial compartments for *in vitro* protein synthesis and product capture. Ribosomes, DNA, the corresponding protein and the reaction product are contained and kept physically separated within the reverse micelles. In the original demonstration of this technology, the enzyme synthesized within the micelle modified the DNA, affording a means for enrichment.

High throughput assays utilizing fluorescence detection are gaining momentum for library screening applications. Two-dimensional fluorescence imaging has been used for identifying clones having a desired fluorescence signature that may then be isolated by micromanipulation (Joo *et al.,* 1999, Zuck *et al.,* 1999). Microfluidic systems for continuous-flow biochemical or cell-based assays are also coming of age (Sundberg, 2000). However, at present, flow cytometry represents the most advanced library screening tool, in terms of instrument sophistication, throughput, and the degree of enrichment of positive cells. A number of recent studies that will be discussed later in this chapter have highlighted the utility of flow cytometry as a tool for library screening applications (see Daugherty *et al.,* 2000a, for practical details on the screening of biosynthetic libraries).

Flow cytometry is a discipline in itself and the relevant details cannot be discussed here. The interested reader should consult the superb text *Practical Flow Cytometry* by Howard Shapiro (Shapiro, 1995). Briefly, flow cytometry is a one-dimensional implementation of fluorescence microscopy, where the particles or cells under observation flow in a stream past a light source and optical collection apparatus. The light scattering characteristics of the observed cell or particle are determined, as is the fluorescence emission due to biological fluorescent molecules and/or synthetic substrates that interact with cellular macromolecules. Flow cytometers are designed to be easily programmed to collect cells of the desired optical properties, such as light scattering or fluorescence profile, using either a closed cycle or an electrostatic droplet deflection technology. Light scatter and/or fluorescence intensity, and usually both, are used to ''gate'' on a specific element of fluid and trigger the sorting mechanism. Electrostatic sorters require that the fluid stream is dispersed into micron-size droplets that are charged and routed into an appropriate container by passing through a set of deflector plates held at a certain potential. Electrostatic sorters are somewhat more expensive, but offer three important advantages: very high sorting rates, more efficient recovery of single cells, and the ability to simultaneously isolate up to four cell subpopulations having different fluorescence or light scattering profiles. The later can be very convenient for protein library screening applications because different populations (e.g., those exhibiting low, moderate, and high activity) can be collected as separate pools for further examination. Close cycle sorting utilizes a mechanical or piezoelectric device that disrupts the flow stream and diverts a volume element of fluid into a container. Mechanical and piezoelectric sorters are generally less expensive, but the maximum sorting rates are about 2000 cells/sec, more than forty times lower than those of top-of-the-line commercial electrostatic sorters. A mechanical, microfluidic sorter was recently demonstrated (Fu *et al.,* 1999). It offers the potential for running multiple devices in parallel, which may ultimately help achieve more satisfactory throughput rates.

The general methodology for the isolation of target cells from libraries is shown in Figure 1 (see color insert). As a rule, cells with a desired fluorescence profile are enriched by factors of 500 to 50,000 per round of sorting. Therefore, multiple rounds of enrichment and regrowth (amplification) of the sorted population are needed to isolate clones that are present at a low frequency in large libraries.[2] The actual enrichment factor that can be achieved in a single round of sorting depends on several parameters including the signal to noise ratio (i.e., the difference between the fluorescence of the target cells compared to background cells expressing proteins with weak or no function), how many parameters can be used to gate precisely on the target cells, and the level of protein expression per cell (Daugherty *et al.,* 2000a).

The screening of protein libraries by flow cytometry offers the following advantages:

1. It is a truly high throughput screening technique. As many as 1×10^9 cells/hr can be processed with state-of-the-art research instrumentation and somewhat lower but comparable rates are attainable with top-of-the-line commercial instruments. The isolation of rare clones represented within a heterogeneous population at frequencies as low as $1:10^7$ has been demonstrated (Leary, 1994).
2. Quantitative assays can be performed on large populations, with single cell resolution. Flow cytometric analysis provides the opportunity to examine the distribution of protein functions within a library and to determine the fraction of clones having an activity of interest.
3. Ligand binding equilibria and dissociation kinetics can be readily determined by flow cytometry of whole cells (or, if desired, using proteins

[2] Alternatively, single cells can be sorted into microtiter well plates. The cells are amplified by adding the appropriate growth media and the resulting clonal populations are screened using conventional liquid assays. Sorting into microtiter well plates circumvents problems resulting from growth competition. A greater degree of enrichment of positive clones is obtained since there is no growth competition among the sorted clones. However, a means for carrying out liquid phase assays in a high throughput manner is necessary.

immobilized on beads). Labeling the cells with a concentration of ligand above the K_D followed by addition of competitor for a specified period of time can be used to select for mutant proteins that exhibit a desired range of rate constants for ligand dissociation. This is a very useful feature for the isolation of high affinity binders. The ligand binding kinetics of clones isolated from a library screening experiment can be directly determined flow cytometrically, eliminating the need for purifying the protein. The kinetics of ligand binding, as determined by flow cytometry, are in good agreement with measurements using other surface techniques such as Surface Plasmon Resonance (Daugherty *et al.,* 1998, Boder *et al.,* 2000).

4. Unlike phage display and other screening technologies that rely on ligand binding, flow cytometry can be readily used to select clones either on the basis of binding or catalytic activity. Numerous enzymatic assays have already been adopted for use with flow cytometry and suitable commercial probes are available from vendors such as Molecular Probes.

5. Multiple quantitative parameters for each cell can be analyzed simultaneously, including various fluorescence signals and forward and 90° light scattering. The ability to carry out multiple assays on each cell is important for minimizing false positives and for the facile selection of mutants having a desired specificity. For example, mutations that give a high signal in a functional assay due to an increase in expression level, rather than a *bona fide* change in specific activity, can be easily detected. In addition, enzymatic activity toward two or more substrates can be detected in one step by monitoring the reaction of the corresponding probe molecules containing dyes that emit at different wavelengths.

Of course, as with any assay technique, flow cytometry is not without limitations. First, a flow cytometric assay for the desired function must be developed. Detection of ligand binding is usually straightforward and simply requires the conjugation of the ligand to the fluorescent dye of choice. On the other hand, there are no fluorescent assays for certain enzymatic reactions. Even when fluorescent assays are available, they may not be useful for flow cytometry because they are not retained by the cell. Substrates that give a bright signal by fluorescence microscopy are often not useful for flow cytometry because the signal to noise ratio is not sufficiently high for rare cell sorting. This is the case when there is significant cell autofluorescence at the emission wavelength of the probe or when there is a high degree of nonspecific adsorption of the probe onto the cells. A related, but less significant, issue is that the excitation wavelength of several useful dyes for enzymatic assays is not compatible with the laser setup of many flow cytometry facilities.

Second, the flow cytometric isolation of very rare clones from a library is technically demanding. For most protein engineering applications microorganisms are used as the expression host. Whereas the sorting of rare mammalian cells is widely practiced in cell biology research and in clinical medicine and is therefore highly advanced, the screening and isolation of microbial cells is still much more of an art (Davey and Kell, 1996).

The analysis of libraries by flow cytometry requires special attention to be paid to the protein expression system. An ideal expression system for isolation of live cells by flow cytometry must fulfill the following criteria (Daugherty *et al.,* 1999):

1. Tight on/off regulation. Protein expression must be strongly repressed when the cells are grown for many generations to maintain proportionate representation of target cells, despite a possible growth advantage for some non-target cells.
2. Under conditions of full induction the level of expression must be sufficient to give a satisfactory fluorescent signal but not so high as to cause cell death. If protein expression has a significant effect on viability, then it may not be possible to use sequential rounds of growth and enrichment for the isolation of very rare cells.
3. A desired level of protein synthesis should be obtainable within a short time after induction. Rapid induction is desirable for two reasons: First, a short induction period reduces the time for each round of screening. Second, and most important, when the cells are kept in an induced state for a shorter period, Darwinian selection of faster growing non-target clones is minimized.

These requirements are relevant when it is important to maintain cell viability and minimize competition among the clones within a sorted pool. The alternative is to sort nonviable cells into microtiter wells and rescue the DNA by PCR amplification. Needless to say, this strategy is more time consuming. However, the sorting of dead cells is necessary when it is desired to select for proteins that are stable under conditions not compatible with cell viability, for example, pH and temperature extremes or the presence of organic solvents.

III. Cell Surface Display Technologies

Ligand binding, and certain enzymatic assays, can be greatly simplified if a protein is biosynthetically anchored on the external surface of a cell. Surface displayed proteins are readily accessible to fluorescent probes and thus, transport limitations are circumvented. In addition, nonspecific labeling due to binding or catalytic turnover of the probe by competing intracellular proteins is eliminated. The molecular architecture of cell surfaces can also be engineered in a variety of ways to facilitate the quantitative capture of small molecules and fluorescent reaction products (Mahal and Bertrozzi, 1997; Olsen *et al.,* 2000).

For a heterologous protein to be anchored on the surface of microorganisms, it has to be compatible with translocation across the lipid bilayer membrane and to contain appropriate signals for secretion, intracellular sorting, and surface anchoring. Vectors for surface expression of heterologous proteins in bacteria were not developed until fairly recently (Georgiou *et al.,* 1997; Stahl and Uhlen, 1997). The first useful system for displaying full-length heterologous proteins on gram-negative bacteria employed a chimeric Lpp-OmpA fusion to target passenger proteins to the cell surface (Francisco *et al.,* 1992). Subsequently, it was shown that flow cytometry could be used to enrich cells displaying scFv antibodies from a 10^5 excess of background cells (Francisco *et al.,* 1993). Since then, Lpp-OmpA fusions have been used to screen libraries of antibodies and other proteins (Christmann *et al.,* 1999; Daugherty *et al.,* 2000b). Other fusion systems for protein display in *E. coli* have been reported in recent years but have not yet been tested for library screening applications (Klauser *et al.,* 1993; Chang *et al.,* 1999; Jung *et al.,* 1999). Techniques for protein display in gram-positive bacteria such as streptococci and staphylococci are also available (Fischetti *et al.,* 1996; Gunneriusson *et al.,* 1996; Stahl and Uhlen, 1997). However, thus far the screening of libraries displayed on gram-positive bacteria has not been reported.

Wittrup and co-workers have successfully used fusions to the *S. cerevisiae* surface protein Aga2 for protein display in yeast (Boder and Wittrup, 1997). This system has proved very useful for the isolation of protein mutants with better expression characteristics and greatly improved ligand binding affinity (Shusta *et al.,* 1999a; Boder *et al.,* 2000). Yeast, being a eukaryotic organism, may be more suitable for the functional expression of some complex proteins (i.e., glycoproteins) that generally fail to fold correctly in bacteria. On the other hand, DNA transformation of *E. coli* is significantly more efficient than that of yeast, making *E. coli* the preferred host for the screening of highly complex libraries.

Finally, protein display on the surface of insect cells has also been described (Ernst *et al.,* 1998). Proteins expressed in insect cells are more likely to be properly glycosylated and correctly folded, relative to microbial hosts. However, constructing protein libraries of significant diversity in insect cells is likely to be a challenge.

IV. Library Screening by Flow Cytometry

A. Ligand Binding

Flow cytometry has so far been used primarily for the screening of protein libraries displayed on the surface of bacteria or yeast. The isolation of high affinity scFv antibodies, single chain T Cell Receptors (scTCRs) and protease inhibitors from libraries screened by flow cytometry has been reported (Boder and Wittrup, 1997; Daugherty *et al.,* 1998; Christmann *et al.,* 1999; Boder *et al.,* 2000; Daugherty *et al.,* 2000b). The following discussion outlines key findings from these studies.

There is evidence that flow cytometric screening can correctly identify the ''best'' clone in a library, that is, the one having the highest ligand binding affinity. Daugherty *et al.* (1998) randomized four CDR 3 residues at or near the V_H–V_L interface of a scFv antibody specific to digoxin. After four rounds of enrichment by flow cytometry, they enriched a clonal population that had the same amino acid sequence as the wild-type residues but was encoded by a different sequence at the DNA level. The frequency of this alternate DNA sequence encoding the wild-type amino acids was only about $1:10^6$. The fact that this very rare clone was isolated to homogeneity by the screen was taken to indicate that, for this set of four residues, the parental amino acid sequence is optimal for digoxin binding. In subsequent studies, Daugherty *et al.* (1998) successfully isolated improved mutants of the anti-digoxin antibody by randomizing V_L residues in the binding pocket Similarly, numerous mutants with subnanomolar affinities were isolated by FACS screening of error-prone PCR libraries of the entire scFv gene (Daugherty *et al.,* 2000b).

Display in *E. coli* and flow cytometry has been used to isolate high affinity mutants of the EETI-II protease inhibitor that bind to trypsin with high affinity (Wentzel *et al.,* 1999). EETI-II is a twenty-eight amino acid protein of the squash family of protease inhibitors. Functional EETI-II could displayed on the surface of *E. coli* indicating correct formation of the three disulfide bonds that form a cystine knot motif. A four amino acid loop in EETI-II was randomized using an NN(G/C) coding scheme and 5×10^7 cells were screened for binding to trypsin. Prior to FACS screening, the library size was reduced eight-fold by a pretreatment step involving binding to biotinylated trypsin followed by capture onto magnetic beads coated with streptavidin. The bacteria eluted from the magnetic beads were sorted by flow cytometry to isolate highly fluorescent cells. Because the particular expression system used for display on the *E. coli* surface resulted in cell lethality, the EETI-II genes in highly fluorescent but nonviable sorted solution had to be rescued by PCR

(Wentzel *et al.,* 1999). Nonetheless, mutants with trypsin affinity similar or greater than the wild type could be isolated, demonstrating that sorting of dead cells followed by PCR rescue of the DNA is useful for library screening purposes. The same research group subsequently showed that cell lethality could be prevented by using an Lpp-OmpA fusion for *E. coli* display, thus enabling the isolation of individual clones without the need for PCR and subcloning (Christmann *et al.,* 1999).

Single chain $V_\alpha V_\beta$ T cell receptors are notoriously difficult to express as soluble proteins in microorganisms. A scTCR from the T cell line 2C was mutagenized *in vivo* using a *mutD E. coli* strain and the resulting library was displayed in *S. cerevisiae.* ScTCR mutants that were compatible with expression in yeast and could be recognized by conformational antibodies specific to either the V_β domain or V_α–V_β interface (Kieke *et al.,* 1999) were isolated. Further analysis showed that the mutants could be produced in soluble form and in appreciable amounts in yeast, whereas the wild-type scTCR protein was retained in the endoplasmic reticulum and could not be secreted from the cell. Both secretion and display on the yeast surface were correlated to the stability of the proteins to chemical denaturation (Shusta *et al.,* 1999b). This important study provides an elegant demonstration of the usefulness of flow cytometric library screening techniques for improving the expression of hard to produce proteins and demonstrates a straightforward route to obtaining large amounts of TCR proteins for structure-function studies. In more recent work Krantz, Wittrup and their co-workers selected scTCR variants with increased binding to fluorescently labeled peptide-MHC from a library in which five CDR residues had been randomized (Holler *et al.,* 2000). Even though a small library (10^5 clones) covering only a portion of the potential sequence diversity was screened, mutants with significantly enhanced affinity (down to the low nanomolar range) and high specificity for a particular peptide-MHC complex were nonetheless selected in a single round (Holler *et al.,* 2000).

Wittrup and co-workers have also carried out extensive studies on the affinity maturation of the 4-4-20 scFv antibody that binds to fluorescein (Boder and Wittrup, 1997). Binding of the hapten occurs with a high affinity (K_D = 0.7 nM in PBS) and results in significant quenching of its fluorescence. Boder and Wittrup (1997) showed that the antifluorescein scFv was displayed on the surface of yeast and that the cells could be fluorescently labeled by FITC-biotin followed by streptavidin conjugated phycoerythrin (PE, a very useful fluorescent protein with a high quantum yield and an extinction coefficient of 2.4×10^6 M^{-1} cm^{-1}). Initially, a mutant with single amino acid substitution (V_HS65A) exhibiting three-fold higher affinity (designated 4M1.1) was isolated from a library generated in an *E. coli* mutator strain (Boder and Wittrup, 1997). In subsequent studies, a library of 4M1.1 random mutants was constructed by error-prone PCR shuffling (Stemmer, 1994a, 1994b). The library was screened by FACS by first incubating with a saturating concentration of labeled ligand, followed by prolonged competition with a large excess of 5-aminofluorescein. Twenty clones were picked at random and used to generate a second library by error-prone PCR shuffling. The new library was then screened under more stringent conditions (i.e., a longer incubation time in the presence of competitor). After four rounds of screening and *in vitro* recombination, the library was incubated in the presence of competitor for several days to ensure that only the slowest dissociating clones retained the labeled ligand and were thus fluorescent. Remarkably, a number of clones exhibiting extremely low rate constants for the dissociation of fluorescein (in the range of $1.2–2.0 \times 10^{-6}$ sec^{-1}) could be isolated. This represents an improvement of four orders of magnitude relative to the parental 4-4-20 scFv antibody. One clone that was characterized in detail was reported to exhibit an equilibrium dissociation constant for fluorescein on the order 270 fM in phosphate buffered saline. The binding of that mutant antibody to fluorescein appears to be the tightest noncovalent protein-ligand binding interaction known (Boder *et al.,* 2000). It is doubtful that any screening technique other than flow cytometry could have yielded such an astonishing degree of functional improvement. Interestingly, the fourth round clones exhibiting subpicomolar affinities had ten consensus amino acid substitutions as well as an additional one to six unique substitutions, depending on the clone. Nine of the consensus amino acid substitutions were in the heavy chain, but only one in the light chain. Nine out of ten consensus mutations in the affinity matured antibodies resulted in amino acid substitutions that are not common among mouse antibodies and occur in less than 10% of the sequences in the Kabat database (Fig. 2, see color insert). One notable example was the nearly invariant V_H103 residue that is Trp in 98% of the antibodies in Kabat, but was replaced with Leu in the affinity improved mutants. This result indicates that the conservation of particular amino acids at certain positions within the immunoglobulin fold could be dictated by biological constraints rather than structural or functional considerations.

B. *Flow Cytometric Analysis of the Effect of the Rate of Random Mutations on Protein Function*

A detailed analysis of the effect of the mutational load on the distribution of protein activity within a mutant population has been carried out for a model scFv antibody that binds to the cardiac glycoside digoxigenin with high affinity (K_D = 2 nM, Daugherty *et al.,* 2000b; R. Loo, G. Chen, B. L. Iverson, and G. Georgiou, unpublished). Briefly, a set of libraries with different mutational loads were constructed by error-prone PCR under a variety of conditions, including the addition of Mn^{2+} ions, biased ratios of dideoxynucleotide triphosphates or doping the reaction with nucleoside analogues (Cadwell and Joyce, 1992; Fromant *et al.,* 1995; Vartanian *et al.,* 1996; Zaccolo *et al.,* 1996). The mutated genes were fused to Lpp-OmpA for display in *E. coli* and libraries consisting 10^6 – 2×10^7 scFv-expressing transformants were obtained. The bacteria were labeled with fluorescent hapten at a concentration ten-fold greater than the apparent K_D of the scFv antibody and also with a viability stain (propidium iodide) to exclude any partially lysed cells that could adsorb the probe nonspecifically. A threshold value for fluorescence was defined such that 95% of all the cells expressing the wild-type antibody have a fluorescence equal to or greater than the threshold. 0.01% of control *E. coli* that did not display the scFv antibody fell within this fluorescence window, indicating that the flow cytometric assay has a dynamic range of about four orders of magnitude. The fraction of library clones expressing mutant scFvs with near wild-type hapten binding is shown in Figure 3. For low mutation rates, the percentage of clones exhibiting fluorescence above the threshold decreased exponentially with the mutational load

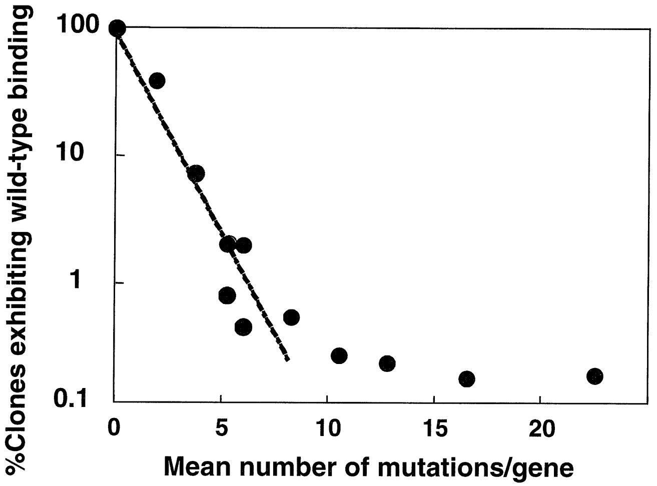


FIG. 3. Percentage of active clones as a function of the mean number of mutations per gene for the 26–10 anti-digoxin single chain Fv antibody. Experimental details can be found in Daugherty *et al.,* 2000b.

($r^2 > 0.95$). However, for libraries with mutation rates in excess of 1% the protein was found to be quite insensitive to the mutational load. Even at a mutation rate of 3% (a mean of 22.5 nucleotide substitutions per gene), 0.18% of all the clones in the library exhibited near wild-type function. Antibody clones exhibiting increased affinity for the digoxigenin hapten were highly represented within the population of active clones. Eight distinct affinity improved mutants expressing antibodies with a K_D two-to four-fold lower than the wild type were isolated from 10^4 clones having fluorescence above the threshold (out of a library of 6×10^6 transformants). The mutants contained between ten and fourteen nucleotide substitutions per gene, a number lower than the mean for the library ($m = 22.5$) but not improbable, given that the distribution of mutations around the population mean is thought to follow Poisson statistics. Gain-of-function mutants having a similar number of nucleotide substitutions and exhibiting the same degree of functional improvement were also isolated from a library with a mean mutation rate of 2% ($m = 14$ nucleotide substitutions). On the other hand, low or no affinity improvement was seen in clones isolated from libraries with mean mutation rates less than 1%. Based on these findings, it is tempting to conclude that for the anti-digoxin scFv antibody (and for libraries of 10^6–2×10^7 random mutants) there is an optimal mutation rate leading to the isolation of functionally improved clones.

In general, only a small fraction of active clones appear to occur in small libraries carrying low rates of mutations (Moore *et al.,* 1997). As the number of mutations is increased, the sequence space expands astronomically. Even with the high throughput afforded by flow cytometry, only a tiny fraction of all possible combinations of five substitutions in a protein of 200 a.q. can be sampled. If most mutations are indeed deleterious, then the probability of finding clones that reain activity, let alone the probability of isolating gain-of-function mutants, by sampling a small portion of the potential diversity would have to be vanishingly small. However, that was not the case with the anti-digoxin antibody (Fig. 3). What is the mechanism responsible for the high percentage of functional proteins found in populations carrying more than twenty nucleotide substitutions per scFv gene? Is this level of sequence plasticity specific to antibodies or is it a property shared by other proteins? How does the frequency of gain-of-function mutants in a population change as a function of the mean mutation rate? With high throughput screening technologies, the analysis of large libraries is now within reach and will surely lead to a better understanding of protein plasticity as well as optimal strategies for directed evolution.

Protein chemists have long suspected that structure and, to a lesser degree, protein function are remarkably robust properties with respect to mutational tolerance (Creighton, 1993). A number of experimental observations support this conclusion, starting with seminal studies on the *lac* repressor spanning a period of almost twenty years (Markiewicz *et al.,* 1994; Suckow *et al.,* 1996). In the *lac* repressor more than 44% of the 338 residues in the protein are tolerant to multiple amino acid substitution and the tolerance to mutation of different residues has been carefully analyzed with respect to the crystal structure. Sauer and co-workers have studied extensively the effect of mutations on Arc, the 53 a.a. homodimeric repressor of P22 (Bowie *et al.,* 1989 and 1990). Nonconservative amino acid substitutions at several positions distributed throughout the polypeptide chain affected stability to various degrees, yet did not seriously perturb the overall protein fold. More recently, it was shown that Arc could tolerate the substitution of fifteen residues with Ala without disruption of the protein conformation (Brown and Sauer, 1999). A mutant with eleven Ala substitutions was shown to not only fold stably but also retain specific DNA-binding activity. It may be argued that Ala substitutions in Arc represent a special case since this protein is primarily α-helical and Ala is well tolerated in α-helices. Nonetheless, the fact that a protein could fold correctly even when 28% of all the residues had been mutated to an amino acid that cannot participate in any significant side chain interactions is striking. An analogous, and perhaps even more remarkable result from the Sauer lab is the recent finding that switching the order of just two amino acids within the sole β-strand in Arc led to an entirely different, thermodynamically stable fold (Cordes *et al.,* 1999).

In other studies, Wain-Hobson and co-workers have selected active variants of dihydrofolate reductase (DHFR, 78 a.a.) from hypermutated libraries (Martinez *et al.,* 1996). Amino acid substitutions were found in all but six residues. Three rounds of mutation and selection led to the isolation of DHFR mutants in which 22% of the amino acids had been substituted. Finally, Palzkill and co-workers have carried out a very systematic mutagenesis of nearly all codons in β-lactamase (Huang *et al.,* 1996). In this case 43 out of 263 residues were found to not tolerate amino acid substitutions. For DHFR and β-lactamase, the two enzymes that have been subjected to systematic mutagenesis of nearly every residue, the fraction of invariable residues was roughly of the same order (9% and 16%, respectively).

One mechanism that may be important in determining protein plasticity is the occurrence of second site suppressor mutations that serve to compensate for deleterious amino acid substitutions that interfere with protein folding. Intragenic second site suppressors are frequently ''global'' in that they can reverse the effect of deleterious mutations scattered throughout the gene. One such global second site suppressor mutation was critical for restoring function in many β-lactamase mutants (Huang and Palzkill, 1997; Huang *et al.,* 1996). The frequency at which second site suppressor mutations were found in a library of random mutants was estimated at 2×10^{-5}. If the frequency of second site suppressors is similar in other proteins, then the occurrence of such compensatory amino acid substitutions may be an important mechanism in mutational tolerance. The picture that emerges from these studies is that amino acid replacements, even nonconservative ones, at the majority of residues in proteins are tolerated. What needs to be examined in more detail is the frequency with which deleterious, compensatory, or beneficial mutations occur in various randomly mutated genes.

C. *Flow Cytometric Screening of Enzyme Libraries*

Flow cytometry is well suited for the analysis of enzyme activity and kinetics at the single cell level (Watson and Dive, 1994). Flow cytometric assays for numerous enzymes including esterases, proteases, peroxidases, lipases, and oxidoreductases[3] are available and are widely used in research and clinical practice. To date, flow cytometry has not been widely exploited as a screening tool for enzyme engineering purposes, but this is rapidly changing.

Flow cytomtric assays require the use of special fluorescent enzymatic substrates. Catalytic turnover leads into the formation of a fluorescent product either due to a covalent modification that results in direct dequenching of the fluoreophore or though the disruption of fluorescence energy transfer (FRET) between donor and acceptor fluorophores. A probe molecule suitable for flow cytometric enzymatic assays must also satisfy the following criteria:

a. The substrate must be capable of permeating into the cell either through passive diffusion or through active transport.

b. The fluorescent product must be retained by the cell and not diffuse away.

c. Both the substrate and the product should not be susceptible to side or competing reactions by unrelated enzymes.

[3] Two excellent sources for information on available enzyme fluorescent probes are the *Handbook of Fluorescent Probes and Research Chemicals* (http://www.probes.com/handbook/sections/0071.html) and the Functional Cell Biochemistry web site (http://www.biochem.mpg.de/valet/valmeth3.html).

d. Finally, the fluorophore should be photostable, have as high a quantum yield as possible, and, if possible, emit at a wavelength at which there is little or no background cell fluorescence.

The most critical and demanding requirement is that the fluorescent product generated by an enzyme is retained quantitatively within the cell. A number of ingenious strategies for product retention have been developed including intracellular precipitation (the Enzyme-Linked Fluorescence or ELF family of fluorescent substrates), attachment of lipophilic anchors, substrates that yield insoluble chromophores, and retention via formation of complexes with intracellular proteins (for examples, see footnote 3 and Zlokarnik *et al.,* 1998).

The design of fluorescent substrates that satisfy both the need for cell permeation and retention of the product is not trivial. One solution to the problem is to exploit expression systems for protein display on the surface of microorganisms. Surface display circumvents the need for substrate internalization and microbial surfaces have a number of unique features that can be exploited for the quantitative retention of fluorescent products. This concept was demonstrated by Olsen *et al.* (2000) (Fig. 4, see color insert). The surface of gram-negative bacteria is highly negatively charged, and FRET substrates with a polycationic tail adsorb to the surface of *E. coli.* Cleavage of the substrate by a surface enzyme disrupts intermolecular quenching, giving rise to a green fluorescent product that is surface retained via the polycationic tail. Cell green fluorescence correlates well with substrate turnover. An additional advantage of this system is that binding of the FRET substrate by mutants having low k_{cat} sequesters the quenching fluororphore within the hydrophobic environment of the binding pocket, thus resulting in red fluorescence emission. As catalysis progresses and more substrate is converted to product, the red fluorescence due to the quencher decreases while the green fluorescence increases (Fig. 5, see color insert). This feature can be convenient for enzyme screening because it allows discrimination between mutants having a threshold of enzymatic activity from mutants exhibiting increased K_m for the substrate. Figure 5 illustrates how the two-color fluorescence profile changes as a function of time for a mutant of the bacterial protease OmpT isolated from a library. This particular mutant was isolated in a screen for second site suppressors of an OmpT variant in which the catalytic triad His had been substituted with Ala. The particular mutant was isolated from a hypermutated library in a single round, had a total of eight amino acid substitutions, and was better able to accommodate bulky side chains in the S2' subsite (Olsen, M., Stephens, D. L., Iverson, B. L., and Georgiou, G. unpublished). In other studies we have screened large libraries consisting of millions of clones for altered catalytic specificity. For example, mutants exhibiting different amino acid preferences in both the S1 and S1′ substites were isolated in a single round of screening of a hypermutated library (Olsen *et al.,* 2000). The mutant proteins contained between six and eight amino acid substitutions, adding further support to the notion that the screening of hypermutated libraries may be a productive strategy for directed evolution.

V. Concluding Remarks

Within the span of less than eight years directed evolution has completely changed our notions of how naturally isolated proteins can be adapted through mutation to assume entirely new functions or stability characteristics (Arnold, 2000). Directed evolution is clearly changing the way we think about proteins (Spiller *et al.,* 1999), and provides an experimental route to studying longstanding notions about how natural protein functional diversity may have arisen (Altamirano *et al.,* 2000). These remarkable successes have been accomplished, to a large degree, through the laborious screening of small libraries. As the ability to analyze and screen very large populations of mutants becomes commonplace, it may be possible to undertake increasingly bolder goals in terms of exploring the limits of structural and functional protein plasticity. The very high throughput and quantitative nature of flow cytometry make it an ideal tool for the exploration of protein sequence space and for the evolution of new function. It is hoped that the examples discussed in this chapter will stimulate further interest in the use of this powerful technology for protein engineering.

Aknowledgments

I am especially grateful to Dane Wittrup for sharing unpublished data and for reading portions of the manuscript. I would like to thank Jennifer A. Maynard, Richard Loo, and Mark Olsen for many useful comments. Flow cytometry research in the author's laboratory has been supported by DARPA, DOE, and the U.S. Army.

References

Altamirano, M. M., Blackburn, J. M., Aguayo, C., and Fersht, A. R. (2000). Directed evolution of new catalytic activity using the α/β-barrel scaffold. *Nature* **403,** 617–622.

Arnold, F. H. (1998). Design by directed evolution. *Acct. Chem. Research* **31,** 125–131.

Arnold, F. H. (2000). Directed enzyme evolution. http://www.che.caltech.edu/groups/fha/Enzyme/directed.html

Atwell, S., and Wells, J. A. (1999). Selection for improved subtiligases by phage display. *Proc. Natl. Acad. Sci. USA* **9,** 9497–9502.

Boder, E. T., and Wittrup, K. D. (1997). Yeast surface display for screening combinatorial polypeptide libraries. *Nature Biotechnol.* **15,** 553–557.

Boder, E. T., Midelfort, K. S., and Wittrup, K. D., (2000). Directed evolution of antibody fragments with monovalent femtomolar antigen-binding affinity. (Submitted).

Bowie, J. U., and Sauer, R. T. (1989). Identifying determinants of folding and activity for a protein of unknown structure. *Proc. Natl. Acad. Sci. USA* **88,** 2152–2156.

Bowie, J. U., Reidhaar-Olson, J. F., Lim, W. A., and Sauer, R. T. (1990). Deciphering the message in protein sequences: tolerance to amino acid substitutions. *Science* **247,** 1306–1310.

Brown, B. M., and Sauer, R. T. (1999). Tolerance of Arc repressor to multiple-alanine substitutions. *Proc. Natl. Acad. Sci USA* **96,** 1983–1988.

Burks, E. A., Chen, G., Georgiou, G., and Iverson, B. (1997). *In vitro* scanning saturation mutagenesis of an antibody binding pocket. *Proc. Natl. Acad. Sci. USA* **94,** 412–417.

Cadwell, R. C., and Joyce, G. F. (1992). Randomization of genes by PCR mutagenesis. *PCR Meth. Applic.* **2,** 28–33.

Chang, H. J., Sheu, S. Y., and Lo, S. J. (1999). Expression of foreign antigens on the surface of *Escherichia coli* by fusion to the outer membrane protein traT. J. *Biomed. Sci.* **6,** 64–70.

Chen, G., Dubrawsky, I., Mendez, P., Georgiou, G., and Iverson, B. L. (1999). *In vitro* scanning saturation mutagenesis of all the specificity determining residues in an antibody binding site. *Prot. Eng.* **12,** 349–356.

Christmann, A., Walter, K., Wentzel, A., Kratzner, R., and Kolmar, H. (1999). The cystine knot of a squash-type protease inhibitor as a structural scaffold for *Escherichia coli* cell surface display of conformationally constrained peptides. *Prot. Eng.* **12,** 797–806.

Cordes, M. H. J., Walsh, N. P., McKnight, C. J., and Sauer, R. T. (1999). Evolution of a protein fold *in vitro. Science* **284,** 325–327.

Creighton, T. E. (1993). *Proteins.* Freeman, New York.

Daugherty, P. S., Chen, G., Olsen, M. J., Iverson, B. L., and Georgiou, G. (1998). Antibody affinity maturation using bacterial surface display. *Prot. Eng.* **11,** 101–108.

Daugherty, P. S., Olsen, M. J., Iverson, B. L., and Georgiou, G. (1999). Development of an optimized expression system for the screening of antibody libraries displayed on the *E. coli.* surface. *Prot. Eng.* **12,** 613–621.

Daugherty, P. S., Iverson, B. L., and Georgiou, G. (2000a). Flow cytometric screening of cell-based libraries. *J. Immunol. Methods.* In press.

Daugherty, P. S., Chen, G., Iverson, B. L., and Georgiou, G. (2000b). Quantitative analysis of the effect of the mutation frequency on the affinity maturation of scFv antibodies. *Proc. Natl. Acad. Sci. USA* **97,** 2029–2034.

Davey, H. M., and Kell, D. B. (1996). Flow cytometry and cell sorting of heterogenous microbial populations: the importance of single-cell analyses. *Microbio. Rev.* **60,** 641–696.

Demartis, S., Huber, A., Viti, F., Lozzi, L., Giovannoni, L., Neri, P., Winter, G., and Neri, D. (1999). A strategy for the isolation of catalytic activities from repertories of enzymes displayed on phage. *J. Mol. Biol.* **286,** 617–633.

Drees, B. L. (1999). Progress and variations in two-hybrid and three-hybrid technologies. *Curr. Opin. in Chem. Biol.* **3,** 64–70.

Ernst, W., Grabherr, R., Wegner, D., Borth, N., Grassauer, A., and Katinger, H. (1998). Baculovirus surface display: construction and screening of a eukaryotic epitope library. *Nuc. Acids Res.* **26,** 1718–23.

Fernandez, P. B. (1999). Technological advance in high-throughput screening. *Curr. Opin. Chem. Biol.* **2,** 597–603.

Firestine, S. M., Salinas, F., Nixon, A. E., Baker, S. J., Smithrud, D., Jordan, D. B., Basarab, G. S. and Benkovic, S. J. (2000). QUEST: A novel bacterial method for selecting enzymes *in vivo. Nature Biotechnol.* **18,** 544–547.

Fischetti, V. A., Medaglini, D. A., and Pozzi, G. (1996). Gram-positive commensal bacteria for mucosal vaccine delivery. *Curr. Opin. Biotechnol.* **7,** 659–666.

Forrer, P., Jung, S., and Plückthun, A. (1999). Beyond binding: using phage display to select for structure, folding and enzymatic activity in proteins. *Curr. Opin. Biotechnol.* **9,** 514–520.

Francisco, J. A., Earhart, C. F., and Georgiou, G. (1992). Transport and anchoring of beta-lactamase to the external surface of Escherichia coli. *Proc. Natl. Acad. Sci. USA* **89,** 2713–2717.

Francisco, J. A., Campbell, R., Iverson, B. L., and Georgiou, G. (1993). Production and fluorescence-activated cell sorting of *Escherichia coli* expressing a functional antibody fragment on the external surface. *Proc. Natl. Acad. Sci. USA* **90,** 10444–10448.

Fromant, M., Blanquet, S., and Plateau, P. (1995). Direct random mutagenesis of gene-sized DNA fragments using polymerase chain reaction. *Analytical Biochemistry* **224,** 347–353.

Fu, A. Y., Spence, C., Scherer, A., Arnold, F. H., and Quake, S. R. (1999). A microfabricated fluorescence-activated cell sorter. *Nature Biotechnol.* **17,** 1109–1111.

Georgiou, G., Stathopoulos, C., Daugherty, P. S., Nayak, A. R., Iverson, B. L., and Curtiss, R. III (1997). Display of heterologous proteins on the surface of microorganisms: from the screening of combinatorial libraries to live recombinant vaccines. *Nature Biotechnol.* **15,** 29–34.

Gunneriusson, E., Samuelson, P., Uhlen, M., Nygren, P. A., and Stahl, S. (1996). Surface display of a functional single-chain Fv antibody on *Staphylococci. J. Bacteriology* **178,** 1341–1346.

Hanes, J., Jermutus, L., Weber-Bornhauser, S., Bosshard, H. R., and Pluckthun, A. (1998). Ribosome display efficiently selects and evolves high-affinity antibodies *in vitro* from immune libraries. *Proc. Natl. Acad. Sci. USA* **95,** 14130–14135.

He, M., and Taussig, M. J., (1997). Antibody-ribosome-mRNA (ARM) complexes as efficient selection particles for *in vitro* display and evolution of antibody combining sites. *Nucleic Acids Res.* **25,** 5132–5134.

Holler, P. D., Holman, P. O., Shusta, E. V., H'Herrin, S., Wittrup, K. D., and Kranz, D. M. (2000). *In vitro* evolution of a T cell receptor with high affinity for Peptide/MHC. *Proc. Natl. Acad. Sci. USA* **97,** 5387–5392.

Huang, W., and Palzkill, T. (1997). A natural polymorphism in beta-lactamase is a global suppressor. *Proc. Natl. Acad. Sci. USA* **94,** 8801–8806.

Huang, W., Petrosino, J., Hirsch, M., Shenkin, P. S., and Palzkill, T. (1996). Amino acid sequence determinants of *β*-Lactamase structure and activity. *J. Mol. Biol.* **258,** 688–703.

Joo, H., Lin, Z., and Arnold, F. H. (1999). Laboratory evolution of peroxide-mediated cytochrome P450 hydroxylation. *Nature* **399,** 670–673.

Jung, H. C., Lebeault, J. M., and Pan, J. G. (1999). Surface display of Zymomonas mobilis levansucrase by using the ice-nucleation protein of Pseudomonas syringae. *Nat. Biotechnol.* **16,** 576–580.

Kieke, M. C., Shusta, E. V., Boder, E. T., Teyton, L., and Wittrup, K. D. (1999). Selection of functional T cell receptor mutants from a yeast surface-display library. *Proc. Natl. Acad. Sci. USA* **96,** 5651–5656.

Klauser, T., Pohlner, J., and Meyer, T. F. (1993). The secretion pathway of IgA protease-type proteins in gram-negative bacteria. *Bioessays* **15,** 799–805.

Koltermann, A., Kettling, U., Bieschke, J., Winkler, T., and Eigen, M. (1998). Rapid assay processing by integration of dual-color fluorescence cross-correlation spectroscopy: high throughput screening for enzyme activity. *Proc. Natl. Acad. Sci. USA* **95,** 1421–1426.

Kuchner, O., and Arnold, F. H. (1997). Directed evolution of enzyme catalysts. *TIBTECH* **15,** 523–530.

Leary, J. F. (1994). Strategies for rare cell detection and isolation. *Methods Cell Biol.* **42,** 331–358.

Loo, R., Chen, G., Iverson, B. L., and Georgiou, G. Unpublished.

MacBeath, G., Kast, P., and Hilvert, D. (1998). Redesigning enzyme topology by directed evolution. *Science* **279,** 1958–1961.

Mahal, L. K., and Bertozzi, C. R. (1997). Engineered cell surfaces: fertile ground for molecular landscaping. *Chemistry & Biology* **4,** 415–422.

Markiewicz, P., Kleina, L. G., Cruz, C., Ehret, S. and Miller, J. H., (1994). Genetic studies of the *lac* repressor XIV. Analysis of 4000 altered *Escherichia coli lac* repressor reveals essential and non-essential residues, as well as 'spacers' which do not require a specific sequence. *J. Mol. Biol.* **240,** 421–433.

Martinez, M. A., Pezo, V., Marlière, P., and Wain-Hobson, S. (1996). Exploring the functional robustness of an enzyme by *in vitro* evolution. *EMBP Journal* **15,** 1203–1210.

Mattheakis, L. C., Bhatt, R. R., and Dower, W. J. (1994). An *in vitro* polysome display system for identifying ligands from very large peptide libraries. *Proc. Natl. Acad. Sci. USA* **91,** 9022–9026.

Moore, J. C., Hua-Ming, J., Kuchner, O., and Arnold, F. H. (1997). Strategies for the *in vitro* evolution of protein function: enzyme evolution by random recombination of improved sequences. *J. Mol. Biol.* **272,** 336–347.

Olsen, M. J., Stephens, D., Griffiths, D., Daugherty, P. S., Georgiou, G., and Iverson, B. L. (2000). Isolation of novel enzymes from large libraries. Submitted.

Pedersen, H., Hölder, S., Sutherlin, D. P., Schwitter, U., King, D. S., and Schultz, P. G. (1998). A method for directed evolution and functional cloning of enzymes. *Proc. Natl. Acad. Sci. USA* **95,** 10523–10528.

Pelletier, J. N., Arndt, K. M. Plückthun, A., and Michnick, S. W. (1999). An in vivo library-versus-library selection of optimized protein-protein interactions. *Nature Biotechnol.* **17,** 683–690.

Roberts, R. W. (1999). Totally *In vitro* protein selection using mRNA-protein fusions and ribosome display. *Curr. Opin. Chem. Biol.* **3,** 268–273.

Roberts, R. W., and Szostak, J. W. (1997). RNA-peptide fusions for the *in vitro* selection of peptides and proteins. *Proc. Natl. Acad. Sci. USA* **94,** 12297–12302.

Rodi, D. J., and Makowski, L. (1999). Phage-display technology-finding a needle in a vast molecular haystack. *Curr. Opin. Biotechnol.* **10,** 87–93.

Sblattero, D., and Bradbury, A. (2000). Exploiting recombination in single bacteria to make large phage antibody libraries. *Nature Biotechnol.* **18,** 75–80.

Shapiro, H. M. (1995). *Practical Flow Cytometry,* 3rd ed. New York, Wiley-Liss.

Shusta, E. V., Kieke, M. C., Park, E., Kranz, D. M., and Wittrup, K. D. (1999a). Yeast polypeptide fusion surface display levels predict thermal stability and soluble secretion efficiency. *J. Mol. Biol.* **292,** 949–956.

Shusta, E. V., VanAntwerp, J., and Wittrup, K. D. (199b). Biosynthetic polypeptide libraries. *Curr. Opin. Biotechnol.* **10,** 117–122.

Spada, S., Krebber, C., and Plückthun, A. (1997). Selectively infective phages (SIP). *Biol. Chem.* **378,** 445–456.

Spiller, B, Gershenson, A., Arnold, F. H., and Stevens, R. C. (1999). A Structural view of evolutionary divergence. *Proc. Natl. Acad. Sci. USA* **96,** 12305–12310.

Stahl, S., and Uhlen, M. (1997). Bacterial surface display: trends and progress. *Trends Biotech.* **15,** 185–192.

Stemmer, W. P. C. (1994a). DNA shuffling by random fragmentation and reassembly: *In vitro* recombination for molecular evolution. *Proc. Natl. Acad. Sci. USA* **91,** 10747–10751.

Stemmer, W. P. C. (1994b). Rapid evolution of a protein *in vitro* by DNA shuffling. *Nature* **370,** 389–391.

Stockwell, B. R., Haggarty, S. J., and Schreiber, S. L. (1999). High-throughput screening of small molecules in miniaturized mammalian cell-based assays involving post-translational modifications. *Chem. & Biol.* **6,** 71–83.

Suckow, J., Markiewicz, P., Kleina, L. G., Miller, J., Kisters-Woike, B., and Müller-Hill, B. (1996). Genetic studies of the lac repressor XV: 4000 single amino acid substitutions and analysis of the resulting phenotypes on the basis of the protein structure. *J. Mol. Biol.* **261,** 509–523.

Sundberg, S. A. (2000). High throughput and ultra-high throughput screening:solution and cell-based approaches'' *Curr Opin. Biotechnol.* **11,** 47–53.

Tawfik, D. S., and Griffiths, A. D. (1998). Man-made cell-like compartments for molecular evolution. *Nat. Biotechnol.* **16,** 652–656.

Vartanian, J-P., Henry, M., and Wain-Hobson, S. (1996). Hypermutagenic PCR involving all four transitions and a sizeable proportion of transversions. *Nucleic Acids Research* **24,** 2627–2631.

Watson, J. V., and Dive, C. (1994). Enzyme Kinetics. *Methods Cell Biol.* **41,** 469–507.

Wentzel, A., Christmann, A., Kratzner, R., and Kolmar, H. (1999). Sequence requirements of the GPNG beta-turn of the *Ecballium elaterium* trypsin inhibitor II explored by combintorial library screening. *J. Biol. Chem.* **274,** 21037–21043.

Winkler, T., Kettling, U., Koltermann, A., and Eigen, M. (1999). Confocal fluorescence coincidence analysis: an approach to ultra high-throughput screening. *Proc. Natl. Acad. Sci. USA* **96,** 1375–1378.

Zaccolo, M., Williams, D. M., Brown, D. M., and Gherardi, E. (1996). An approach to random mutagenesis of DNA using mixtures of triphosphate derivatives of nucleoside analogues. *J. Mol. Biol.* **255,** 589–603.

Zlokarnik, G., Negulescu, P. A., Knapp, T. E., Mere, L., Burres, N., Feng, L., Whitney, M., Roemer, K., and Tsien, R. Y. (1998). Quantitation of transcription and clonal selection of single living cells with beta-lactamase as reporter. *Science* **279,** 84–88.

Zuck, P., Lao, Z., Skwish, S., Glickman, J. F., Yang, K. Burbaum, J., and Inglese, J. (1999). Ligand-receptor binding measured by laser scanning imaging. *Proc. Natl. Acad. Sci. USA* **96,** 11122–11127.

This Page Intentionally Left Blank

FROM CATALYTIC ASYMMETRIC SYNTHESIS TO THE TRANSCRIPTIONAL REGULATION OF GENES: *IN VIVO* AND *IN VITRO* EVOLUTION OF PROTEINS

By CARLOS F. BARBAS III, CHRISTOPH RADER, DAVID J. SEGAL,
BENJAMIN LIST, and JAMES M. TURNER

Department of Molecular Biology, The Scripps Research Institute, La Jolla, California 92037

I. Introduction and Scope of Review

Simple binding is a molecular phenotype that is key in the creation of communities of molecules and life. Binding interactions range in complexity from noncovalent, nonspecific interactions governed by solvent effects to dynamic covalent interactions with molecules along a reaction coordinate of a chemical reaction. This review focuses on our laboratory's use of molecular evolution and selection to shape novel proteins of defined function. Although the types of molecular function we are interested in are quite diverse, they share common roots in molecular recognition. The types of function we will consider in this review involve antibody/antigen interactions of the classical type (simple binding), the nonclassical type (catalysis), and the molecular recognition

0065-3233/00 $35.00

of long DNA addresses in complex genomes. Our approach toward understanding nature's proteins is founded on the premise that to understand them we must be able to create them. Although at first this might appear to be a daunting and perhaps impossible task given that nature's proteins and enzymes have been perfected over eons of evolution, one need only look to the mammalian immune system for encouragement. In the immune system one finds a form of Darwinian evolution that operates on a time scale of weeks rather than eons. Antibodies that bind with high affinity to any given antigen or molecular shape can be quickly created. To do this the immune systems uses a variety of methods such as V-gene shuffling, recombination mechanisms that can create randomized gene segments, and somatic hypermutation that places point mutations throughout the binding site. Perhaps one of the first lessons the immune system provides the protein chemist is the concept of a combinatorial library and the power of looking through large libraries of molecules for the few molecules that exhibit improved functional characteristics. Antibody production in humans results first from the development of a large library of antibodies that approximates the number of B cells in the individual—a few billion. This naive antibody diversity is crafted through recombination and gene shuffling from only about 200 germline encoded genes, V, D, and J. By coupling antibody/antigen recognition with the replication and expansion of these cells, antibodies with the best affinity for the antigen at hand are readily amplified. Further, during the course of amplification or clonal expansion, additional diversification of the antibodies occurs, resulting in the production of a secondary library of proteins from which fitter antibodies can be selected. This iterative process leads rapidly to antibodies required to maintain our health in an otherwise pathogen-filled environment. Such a rapid and successful system of protein evolution is more readily apprehended than the direct study of some of evolution's other products that result from eons of selection. Significantly, not only does the immune system provide insights into evolutionary strategies, it offers an experimental means of testing them.

II. Selecting and Evolving Therapeutic Human Antibodies

Our initial studies concerning directed molecular evolution of proteins centered on the potential to recreate the mammalian immune system *in vitro*. The ability to engineer antibodies has not only facilitated the generation of antibodies to virtually any antigen of interest but also their improvement in terms of binding affinity, specificity, and immunogenicity. In particular, selection and evolution of antibodies using phage display have been key tools in antibody engineering for the past decade.

A. *The pComb3 System*

Our phage display strategies for antibody selection and evolution are based on the pComb3 system that was introduced by Barbas *et al.* in 1991 and demonstrated the first successful display of the 50 kDa Fab fragment of antibodies on the phage surface. Figure 1 shows the latest version of the pComb3 phagemid series, phagemid pComb3X. A detailed description of the pComb3 system including the generation and selection of antibody libraries is now available as a laboratory manual from Cold Spring Harbor Laboratory Press (Barbas *et al.,* 2000). The pComb3 system facilitates the monovalent display of protein libraries on the phage surface. Affinity-based selections are best performed using monovalent phage display to avoid avidity affects that result from the display of multiple copies of the protein. Using the pComb3 system we and others have selected and evolved antibodies and other proteins with high specificity and affinity (Rader and Barbas, 1997).

B. *Synthetic Antibodies*

To explore the potential of antibodies generated outside an animal source, we sought to create them synthetically using insights derived from nature's mechanisms. The *in vitro* production of synthetic antibodies would rely on our increasingly sophisticated understanding of how the structure and genetics of antibodies facilitate their rapid evolution *in vivo.* If these insights were indeed faithful, we should then be able to use them to create novel antibody molecules. Let's first consider the structure of the antibody binding site. Three complementarity determining regions, CDRs, provided by each heavy (H) and light (L) chain protein act together by heterodimerization to form the antibody binding site (Padlan, 1994). The binding site therefore results from the convergence of six hypervariable peptide loops or CDRs. It is primarily the variation in amino-acid sequence in these regions that produces antibodies of differing specificities, that is, antibodies that bind different antigens. *In vivo,* diversity in antibody binding sites is initially encoded in the germline by multiple variable (V), diversity (D), and joining (J) gene segments. CDRs 1 and 2 of H and L chains are encoded within the V regions. Light chain CDR 3 (LCDR3) is produced by the genetic recombination of V and J regions, whereas heavy chain CDR3 (HCDR3) is formed by the recombination of V, D, and J regions. The mechanisms for the fusions of the gene segments are complex and in the case of HCDR3 have the potential to generate more than 10^{14} peptides in this region that differ both in length and sequence (Sanz, 1991). The diversity in the length of the expressed peptide in the HCDR3 produced by these

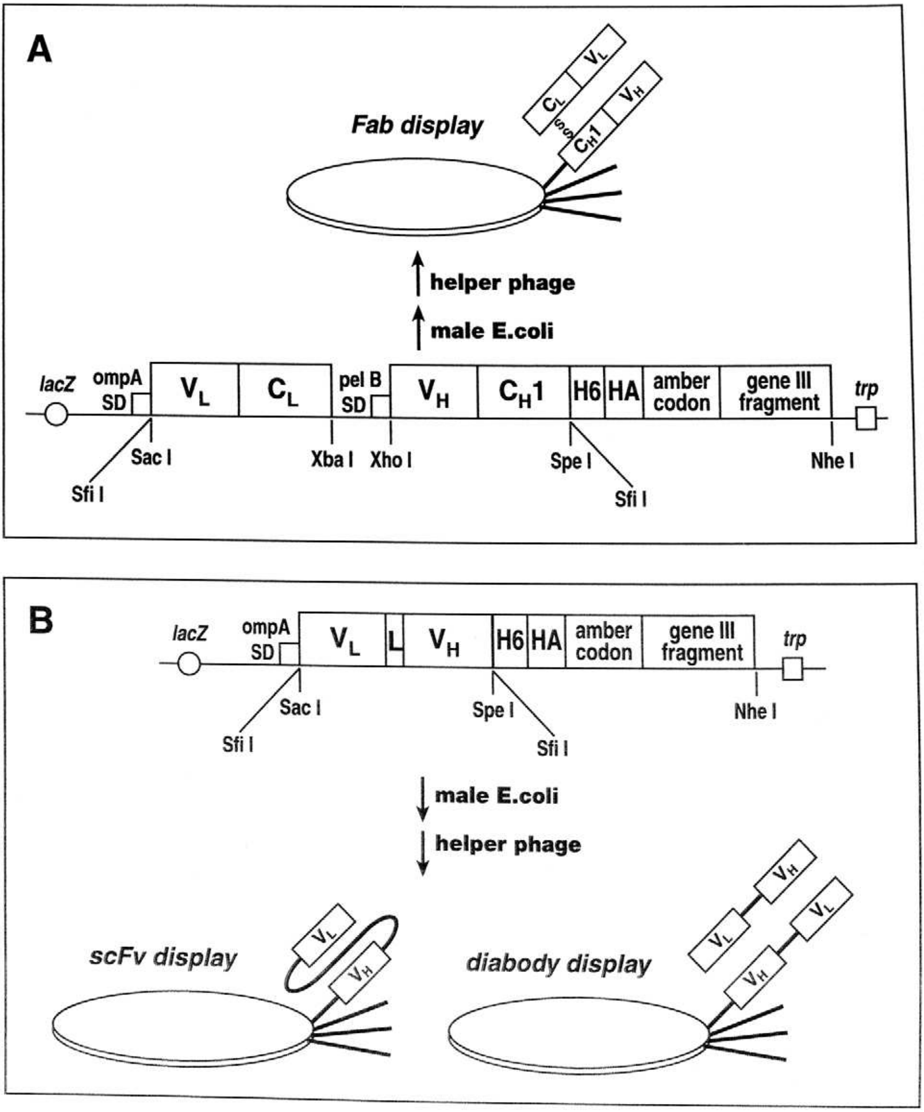


FIG. 1. Phagemid pComb 3X for phage display of antibody libraries. Phagemid pComb3X is the latest version of the pComb 3 phagemid series and an update of phagemid pComb 3H (Rader and Barbas, 1997). Antibody fragments in Fab or scFv format are fused to the C-terminal domain of the minor coat protein of filamentous phage, the gene III protein, which is displayed in low copy number at one end of the phage. Native gene III protein has to be provided to allow infection of *E. coli*. In phagemid systems, helper phage superinfection of *E. coli* transformed with a phagemid leads to the expression of the native gene III protein while the phagemid drives the expression of the fusion protein. Native gene III protein and fusion protein compete in phage assembly. Minimizing the phagemid-driven expression of the fusion protein provides mostly phage that bear either no or just a single copy of the displayed protein. This effectively allows for

rearranged segments is astounding in structural terms, as HCDR3 length may vary from two to more than twenty-six amino acid residues (Wu *et al.,* 1993). In a sense, this CDR segment can be viewed as a random sequence segment of variable length. From structural studies it has become apparent that despite diversity in sequence exhibited in the V-gene segments, the structures of these segments are relatively constrained (Al-Lazikani *et al.,* 1997). Only HCDR3 is not restricted to a small repertoire of main chain conformations or canonical structures (Morea *et al.,* 1998). Given the limited diversity encoded in the germline, a key feature for the generation of new antibody specificities should be HCDR3.

To test this hypothesis, we created the first synthetic antibody libraries in 1992 (Barbas *et al.,* 1992). In this study a repertoire of synthetic

monovalent phage display. (A) In the Fab display version of pComb 3X a single lacZ promoter drives the synthesis of a dicistronic transcript. Two ribosome-binding sites (Shine and Dalgarno sequences; SD) give rise to the translation of two separate polypeptide chains. One encodes a complete light chain (V_L-C_L), the other the Fd fragment of a heavy chain (V_H-C_H1) fused to two peptide tag encoding sequences and the C-terminal domain of the gene III protein. The two peptide tags are a hexa-histidine (H6) peptide and a hemagglutinin decapeptide (HA). At the junction of the peptide tag encoding sequences and the gene III fragment is an amber stop codon, TAG. The gene III fragment is followed by the trp transcriptional terminator. To generate a combinatorial antibody library, light chain and Fd fragment encoding sequences are amplified independently and combined randomly in pComb 3X. The Sac I and Xba I sites are provided for cloning light chain encoding sequences. The Xho I and Spe I sites are for cloning Fd fragment encoding sequences. Alternatively, the sequences encoding light chain and Fd fragment can be fused by overlap extension PCR and cloned directionally using two asymmetric sites of the rare eight-base cutter Sfi I. Sfi I sites are virtually never found in immunoglobulin sequences, making Sfi I a versatile cloning enzyme for the generation of antibody libraries with minimal biases. The leader peptides ompA and pelB target both polypeptides to the periplasm of *E. coli,* which the soluble light chain and the membrane-bound Fd fragment associate. The transformation of male *E. coli sup* strains (for amber stop codon suppression) that overexpress the lacI repressor (for minimizing lacZ driven expression) with Fab encoding pComb 3X followed by superinfection with helper phage leads to the production of phage that, as their phenotype, display one copy of Fab fragment linked to the phage surface by the gene III protein fragment and that contain the corresponding single-stranded phagemid as their genotype. (B) In the scFv display version of pComb 3X the variable immunoglobulin domains of light chain and heavy chain, V_L and V_H, are fused by a peptide linker (L). While a long, flexible linker allows for the monovalent scFv display, a shorter linker leads to the formation of scFv dimers with two antigen binding sites (diabodies). Diabody display requires the recruitment of a soluble scFv released by proteolysis from its gene III protein fusion partner in the periplasm. Following selection, soluble Fab, scFv, or diabodies can be produced in an *E. coli* non-*sup* strain (amber stop codon is not suppressed) after induction with isopropyl β-D-thiogalactopyranoside. The H6 and HA peptide tags facilitate detection and purification of the soluble antibody fragments. Soluble antibody fragments without peptide tags can be produced by digestion of pComb 3X, encoding a selected Fab or scFv, with the compatible restriction enzymes Spe I and Nhe I and self-ligation.

random HCDR3 sequences was constructed on the backbone of an existing human antibody of defined specificity, an anti-tetanus toxoid antibody, thereby fixing the V-gene component of diversity. The sixteen amino-acid HCDR3 region of this antibody was then randomized using the polymerase chain reaction and synthetic oligonucleotides. Typically, segmental mutagenesis of this type is performed using oligonucleotides synthesized with degeneracies at the position at which randomization is desired. The most common oligonucleotide doping strategies use NNK or NNS codon equivalents. N represents an equal mixture of the 4 possible nucleotides A, C, G, and T. K is a mixture of just G and T and S is a mixture of only G and C. Both NNK and NNS doping strategies encode all twenty amino acids and a single amber stop codon within a total of thirty-two codons. Thus, each amino acid is not equivalently represented in the mixture and the biases present in nature's sixty-four codon genetic code remain. To generate novel binding specificities from the single anti-tetanus toxoid binding antibody used to construct this first synthetic antibody library, randomization over the antibody's extended HCDR3 length ensured structural diversity even though such a library would be incomplete as more than 10^{24} clones would be required for each possible amino-acid sequence to be represented. The synthetic genes were cloned to yield a library of approximately 10^{8} clones, a size of the order of the number of B cells in a mouse. Selection of this synthetic repertoire for binding fluorescein resulted in antibodies with affinities in the 20 nM range for the fluorescein-BSA conjugate. Thus, randomization of this single CDR segment allowed for remodeling of an antibody originally evolved to bind a protein to now bind a hapten for which the parental antibody had no measurable affinity. If this change in specificity is expressed in terms of affinity improvement, then a $>10^{4}$-fold improvement in affinity for this antigen was achieved. The use of a single antibody template in these studies was essential in establishing the link between the selected protein and the initial template antibody. Use of a library of antibodies on which HCDR3 diversity would be built would not have allowed the key nature of HCDR3 mutations in hapten binding to be revealed or the hypothesis concerning the central role of HDCR3 to be tested. An interesting result of these studies was the selection of an aspartate residue at position 101 in the HCDR3. In naturally occurring antibodies Asp-101 has been shown to participate in a structurally important salt bridge with Arg-94, the last residue in framework 3. Thus, these synthetic antibodies recapitulated a structurally conserved feature of natural antibodies and support the hypothesis for the role of this conserved sequence. The amino-acid distribution within the HCDR3 of natural antibodies is not completely random and examination of selected synthetic repertoires also reflects this fact. The most conserved portion of HCDR3 is the carboxy terminal region that is contributed by the J-gene segment. This region is characteristically rich in aromatics and, in most cases, encodes Asp at position 101. This distribution is also clearly seen in selected synthetic repertoires. Indeed, sequence similarities between natural and synthetic antibodies can be close enough to allow for the accurate prediction of cross-reactivity for some synthetic antibodies (Barbas and Wagner, 1995). Extension of this approach to accommodate additional diversity in the V-gene segments and CDR regions has also been reported and includes the use of designed consensus protein frameworks (Knappik *et al.,* 2000). It has been shown that this approach allows for the isolation of monoclonal antibodies with specificity for virtually any given antigen (Barbas 1995; Pini *et al.,* 1998).

A more comprehensive discussion of the *de novo* creation of human antibodies from naive and synthetic antibody repertoires can be found in Rader and Barbas (2000). Libraries prepared from immune humans as well as those more rationally designed to interact with a specific receptor have also been explored. Among the types of antibodies prepared are those that target the envelope glycoprotein gp120 of HIV-1 as well as other viruses (Barbas *et al.,* 1992; 1993a; 1994; Burton *et al.,* 1994; Yang *et al.,* 1995) and angiogenic targets such as the human integrin $\alpha_v\beta_3$ (Barbas *et al.,* 1993b; Smith *et al.,* 1994). In addition to providing a rich source of therapeutic antibodies, immune libraries, such as those derived from natural viral infections, allow the interplay between the host immune system and viruses to be probed for vaccine design. Studies concerning immune libraries fall outside the scope of this review, which is more concerned with methods to change template proteins derived from nature rather than their isolation from a natural source. For a review of studies concerning immune libraries, the reader is directed to Burton and Barbas (1994).

C. *Antibody Humanization*

Immune libraries and evolutionary selection strategies intersect in the area of antibody humanization. Recently, we have extended the repertoire of methods available for the generation of therapeutic human antibodies by developing phage display strategies for the selection and humanization of antibodies from immune animals other than mice. Our aim here was to modify protein sequence, in some cases in a very radical way, and yet retain the function of the parental antibody. In the following discussion, we will focus on these novel approaches.

In light of new technologies that allow the selection of immune, naive, and synthetic human antibody libraries displayed on phage as well as from transgenic mice containing human immunoglobulin loci (for recent reviews cf. Hudson, 1999; Vaughan *et al.,* 1998; Rader and Barbas, 1997), the question arises whether the humanization of antibodies from immune animals is outdated. In fact, nonhuman antibodies are highly immunogenic in humans, which severely limits their clinical applications, making antibody humanization mandatory if repeated administration is required for therapy. However, immune animals are still an important and ready source of therapeutic human antibodies for the following reasons:

1. More than twenty years of rodent monoclonal antibody generation by the classical hybridoma technology has yielded a number of promising therapeutic candidates and their humanization compares well to the *de novo* generation and characterization of human antibodies for accessing clinical applications in the coming years.

2. Phage display allows the generation of monoclonal antibodies from virtually any species whose immunoglobulin genes are sequenced. The ability to generate monoclonal antibodies from a variety of species will extend the accessible epitope repertoire of a given antigen. For example, epitopes that are not immunogenic in mice, a species from which the vast majority of monoclonal antibodies to human antigens have been generated, might be immunogenic in rabbits. This is of particular interest for the development of therapeutic human antibodies, that are evaluated in mouse models of human disease in which antibodies are required to recognize both the human antigen and its mouse homologue. In contrast, human antibodies that are derived from immune mice, either indirectly through humanization or directly through transgenic mice containing human immunoglobulin loci, are negatively selected against epitopes displayed by the mouse homologue. Using phage display, monoclonal antibodies to highly conserved human antigens that are not immunogenic in mammalian species can be generated in a nonmammalian species, such as chickens. Detailed protocols for the generation and selection of antibody libraries from immune mice, rabbits, and chickens can be found in Barbas *et al.* (2000).

3. In contrast to human antibodies derived from large naive or synthetic human antibody libraries, antibodies from immune animals were subjected to *in vivo* selection and are therefore more likely to recognize a given antigen selectively (i.e., without cross-reactivity to another antigen).

4. Although until recently antibody humanization was based on arduous rational design strategies that require computer modeling and itera-

tive optimization, selective approaches based on phage display have facilitated antibody humanization.

CDR grafting is the most frequently used rational design strategy for antibody humanization (Riechmann *et al.,* 1988). In this approach, the six CDR loops composing the antigen binding site of the nonhuman antibody are grafted onto corresponding human framework regions. CDR grafting takes advantage of the conserved structure of the variable immunoglobulin domains, with the four framework regions serving as a scaffold that supports the CDR loops. CDR grafting often yields humanized antibodies with much lower affinity because framework residues are involved in antigen binding, either indirectly, by supporting the conformation of the CDR loops, or directly, by contacting the antigen (Foote and Winter, 1992). Therefore, in addition to CDR grafting, it is usually necessary to replace certain framework residues. The fact that about thirty framework residues potentially contribute to antigen binding (Foote and Winter, 1992) makes this fine-tuning step very laborious.

We used a selective approach that combines CDR grafting with framework fine tuning by phage display (Rosok *et al.,* 1996; Baca *et al.,* 1997) for the first reported humanization of a rabbit antibody (Rader *et al.,* 2000). As outlined in Figure 2, this strategy involves the diversification of a small number of framework residues to allow the selection of either the human or the original nonhuman residue. The diversified framework residues are chosen out of a set of key framework residues that are known to be involved in antigen binding. Key framework residues have been identified structurally by crystallography and molecular modeling as well as empirically by antibody humanization. We applied this strategy to the humanization of a rabbit antibody directed to human A33 antigen (Rader *et al.,* 2000). The rabbit antibody had been selected by phage display from antibody libraries generated from immune rabbits. Human A33 antigen is a tumor antigen expressed on the surface of colon cancer cells. The humanized antibodies were found to retain both high specificity and affinity for human A33 antigen (Table I).

Where as the latter approach involves a very limited sequence diversification, another selective approach for antibody humanization diversifies entire V genes. This strategy (Fig. 3) preserves only the original nonhuman CDR3 sequences of light and heavy chain while subjecting the remaining sequence to selection from naive human V-gene libraries. Using designed combinatorial V-gene libraries, we humanized mouse monoclonal antibody LM609 (Rader *et al.,* 1998). LM609 is directed to human integrin $\alpha_v\beta_3$ and has potential applicability in cancer therapy as an anti-angiogenic agent. We demonstrated that this approach (1) provides a rapid route for antibody humanization constraining the

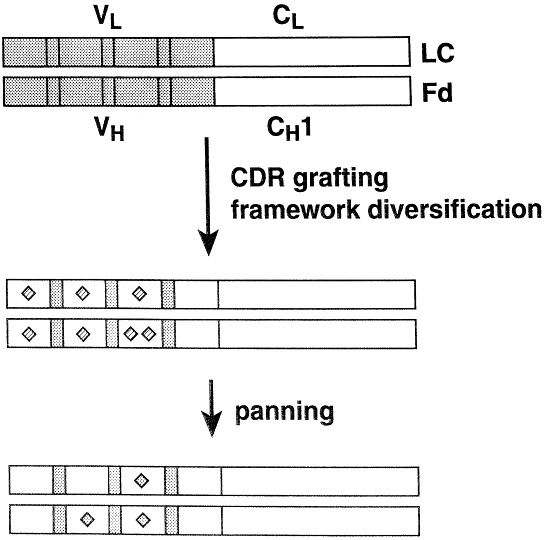


FIG. 2. Selection strategy for antibody humanization that combines CDR grafting with framework fine-tuning. Nonhuman sequences are shown in gray, human sequences are white. The six CDRs of the nonhuman Fab are grafted into a human framework that contains a set of diversified residues. These residues are diversified to allow the selection of either the human or the original nonhuman residue. Reprinted with permission from Rader and Barbas (2000).

TABLE I
Binding Parameters of Rabbit and Humanized Fab Directed to Human A33 Antigen[a,b]

Clone	$k_{on}/10^4$ (M^{-1} s^{-1})	$k_{off}/10^{-4}$ (s^{-1})	K_d (nM)
Rabbit			
1	30.7	1.2	0.39
2	17.4	2.8	1.6
Humanized			
A	10.5	5.9	5.6
B	35.2	6.1	1.7
C	10.9	19.7	18.1
D	6.5	19.0	29.2
E	6.5	5.0	7.7
F	19.2	6.8	3.5

[a] From Rader *et al.* (2000).

[b] Association (k_{on}) and dissociation (k_{off}) rate constants were determined using surface plasmon resonance. Human A33 antigen was immobilized on the sensor chip. K_d was calculated from k_{off}/k_{on}. Clones **A–F** represent six humanized versions of rabbit clone **2.**

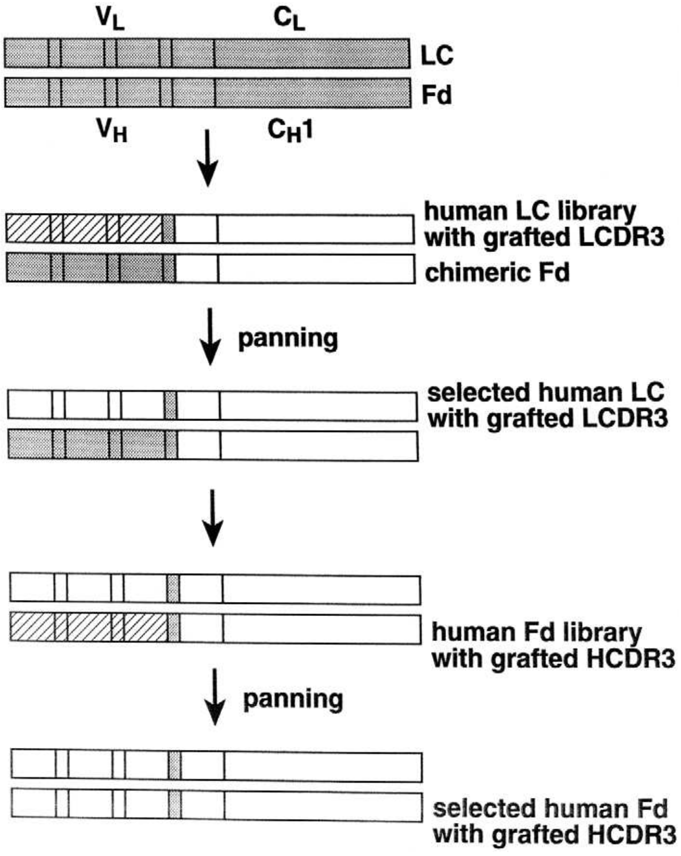

V_L
C_L
LC
Fd
V_H
C_H1
human LC library with grafted LCDR3
chimeric Fd
panning
selected human LC with grafted LCDR3
human Fd library with grafted HCDR3
panning
selected human Fd with grafted HCDR3

FIG. 3. Antibody humanization using designed combinatorial V-gene libraries. Nonhuman sequences are shown in grey, human sequences are white. This strategy involves two selection steps for the sequential humanization of the light chain and the Fd fragment of the heavy chain. Throughout these selections, the only preserved sequences in the variable domains of light chain (V_L) and heavy chain (V_H) are the CDR3 sequences. In the first step a chimeric nonhuman/human Fd fragment is used as a template for the selection of a human light chain that contains the grafted CDR3 loop of the original nonhuman light chain. In the second step a human Fd fragment that contains the grafted CDR3 loop of the original nonhuman Fd fragment is selected. Reprinted with permission from Rader and Barbas (2000).

content of original mouse sequences in the final antibodies to the most hypervariable of the CDRs, (2) generates several humanized versions with different sequences at the same time, (3) results in affinities as high as or higher than the affinity of the original antibody (Table II), and (4) retains the antigen and epitope specificity of the original antibody.

TABLE II
Binding Parameters of Mouse and Humanized Fab Directed to Human Integrin $\alpha_v\beta_3$[a,b]

Clone	$k_{on}/10^4$ (M^{-1} s^{-1})	$k_{off}/10^{-4}$ (s^{-1})	K_d (nM)
Mouse*	14	4.6	3.3
Mouse	8.6	8.6	10
Humanized			
11	1.0	16	160
7	18	5.4	3.0
4	6.8	5.8	8.5
24	13	9.9	7.6
2	11	7.5	6.8

[a] From Rader *et al.* (1998).

[b] Binding kinetics were determined using surface plasmon resonance. Human integrin $\alpha_v\beta_3$ was immobilized on the sensor chip. K_d was calculated from k_{off}/k_{on}. Clones **11, 7, 4, 24,** and **2** represent five humanized LM609 versions. Fab were produced by *E. coli* except for mouse* which was prepared from IgG by papain digestion.

In addition to mouse LM609, we recently humanized a rabbit antibody using designed combinatorial V-gene libraries. Rabbit antibody ST6 specifically binds human chemokine receptor CCR5 that, together with CD4, is required for cell entry of macrophage-tropic HIV-1 viruses. The intracellular expression of a single-chain Fv fragment of ST6 with a C-terminal fusion to the peptide KDEL for retention in the endoplasmic reticulum efficiently blocked surface expression of CCR5 (Steinberger *et al.,* 2000). As shown in Figure 4, a $CCR5^+/CD4^+$ human lymphocyte cell line expressing intrabody ST6 was protected from infection by macrophage-tropic HIV-1 viruses. ST6 was originally selected by phage display from antibody libraries generated from immune rabbits. Interestingly, the humanization of ST6 (Steinberger *et al.,* in preparation) was found to require conservation of CDR2 of the heavy chain in addition to CDR3 of light and heavy chain. This finding can be explained with an important role of HCDR2 of ST6 in antigen binding. The human heavy chain repertoire might not contain CDR2 sequences that are similar enough to support specific binding to the antigen. Containing three grafted CDRs, humanized ST6 was found to retain both high specificity and affinity for human chemokine receptor CCR5.

D. CDR Walking

As described for antibody humanization, phage display strategies for selecting new antibodies can also be utilized to evolve existing antibodies.

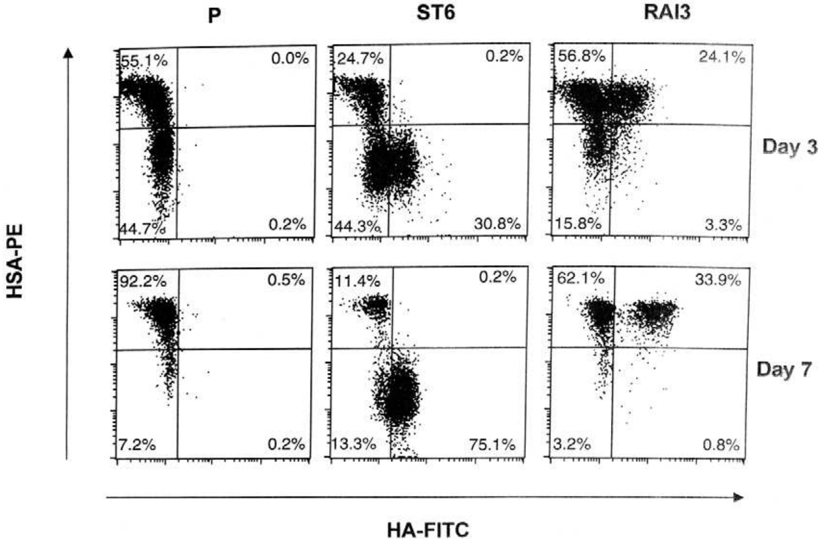


FIG. 4. Cells expressing intrabody ST6 are resistant to HIV-1 infection. The CCR5$^+$/CD4$^+$ human lymphocyte cell line PM1 infected with a macrophage-tropic HIV-1 reporter virus was cocultured with the parental PM1 cell line (P; left), an ST6 intrabody expressing PM1 cell line (ST6; center), and an irrelevant intrabody expressing PM1 cell line for negative control (RAI3; right). Double staining of cells for intrabody expression (HA-FITC; *x* axis) and reporter virus infection (HSA-PE; *y* axis) after three and seven days of coculturing is shown. Note that while the number of infected cells (HSA$^+$) expressing an irrelevant intrabody (HA$^+$) increased from 24% on day 3 to 34% on day 7, virtually all ST6 intrabody expressing cells remained uninfected (Steinberger *et al.,* 2000).

Phage display provides a format for the directed evolution of antibody affinity (Barbas and Burton, 1996; Rader and Barbas, 1997). A detailed description of the strategies that have been developed for the affinity maturation of antibodies can be found in Rader and Barbas (2000). A strategy known as CDR walking (Barbas *et al.,* 1994) stands out as the most general approach for affinity maturation of antibodies. As outlined in Figure 5, CDR walking involves the sequential or parallel optimization of CDRs by randomization of key recognition sequences and subsequent selection by phage display. A key feature of this approach is that mutations are targeted to the CDR regions, also known as the hypervariable regions of antibodies. The reasons for targeting these regions are that (1) sequence mutations in these regions are less likely to create an immunogenic antibody and (2) they are key regions involved in direct and indirect interaction with antigens. Although approaches such as error-prone PCR or DNA shuffling would also be expected to provide diversification for the evolution of antibody affinity, the mutations that

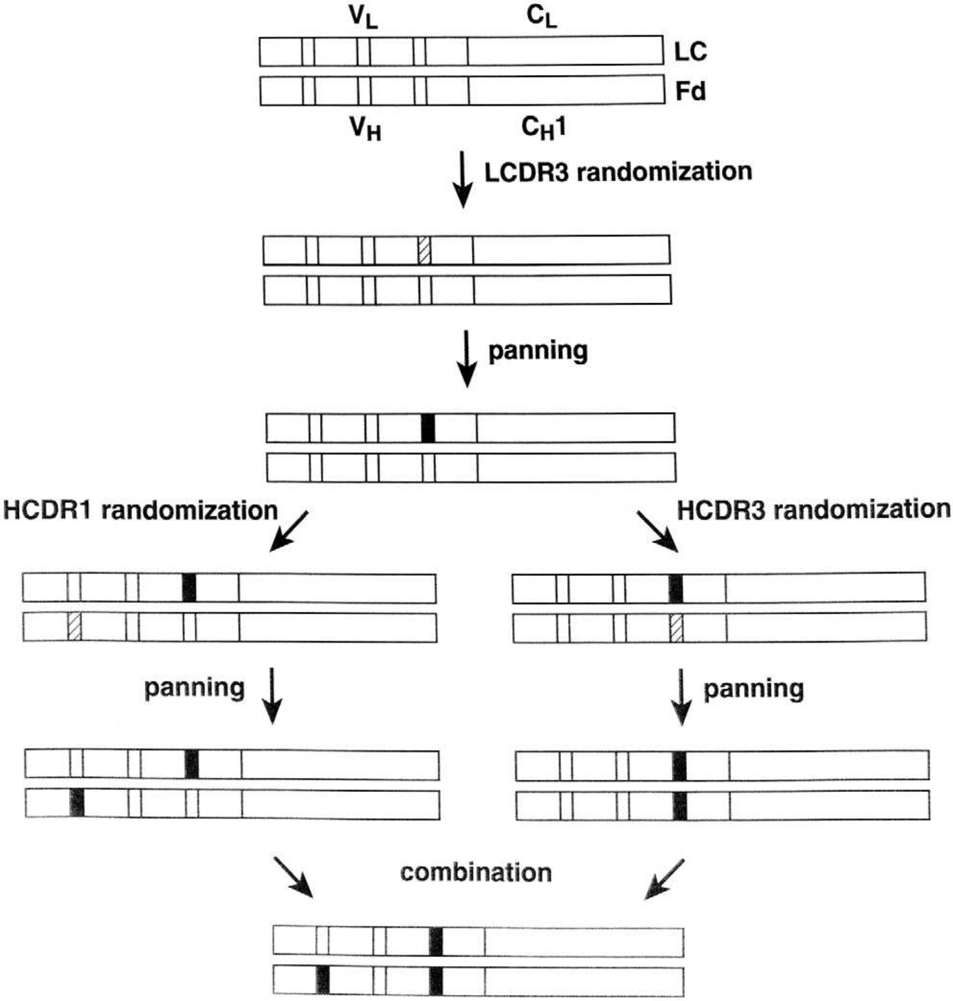


FIG. 5. Affinity maturation of antibodies by CDR walking. CDR walking involves the sequential or parallel optimization of CDRs by randomization of key recognition sequences and subsequent selection by phage display. Shown is a combination of sequential and parallel CDR optimization that was used in the affinity maturation of humanized LM609 (Rader *et al.,* in preparation). CDR3 of the light chain was optimized first, followed by a parallel optimization of CDR1 and CDR3 of the heavy chain. Optimized CDR1 and CDR3 of the heavy chain were combined in the last step.

they would introduce into the antibody would likely be immunogenic if the mutations fell outside the CDR regions. Using CDR walking strategy, we have improved the monovalent affinity of a human antibody to human envelope glycoprotein gp120 of HIV-1 from the nanomolar to picomolar range (Yang *et al.,* 1995). Other groups have reported similar success in the affinity maturation of antibodies by CDR walking (Schier *et al.,* 1996).

III. Directing Evolution at the Level of Chemical Mechanism: Aldolase Antibodies and Asymmetric Catalysis

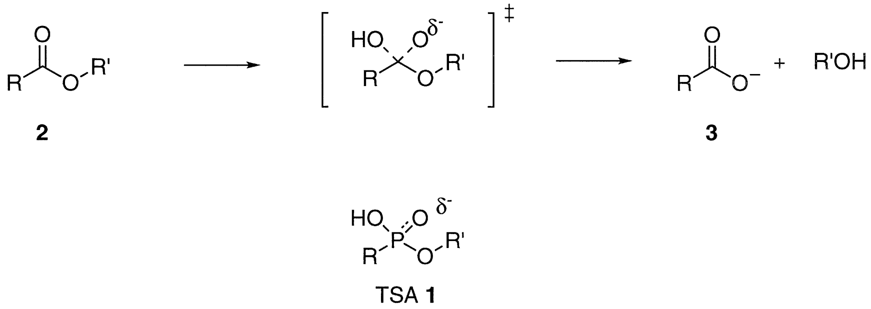
A. Selecting Antibody Enzymes

1. Transition State Analogs

Traditionally, antibody enzymes have been obtained by immunizing with chemically inert transition state analogs (TSAs). This was based on Linus Pauling's notion that enzymes provide rate acceleration for chemical reactions by binding transition states (Pauling, 1946, 1947). Later Bill Jencks restated Pauling's insight concerning catalytic antibodies more precisely (Jencks, 1969): ''If complementarity between the active site and the transition state contributes significantly to enzymatic catalysis, it should be possible to synthesize an enzyme by constructing such an active site. One way to do this is to prepare an antibody to a haptenic group which resembles the transition state of a given reaction. The combining sites of such antibodies should be complementary to the transition state and should cause an acceleration by forcing bound substrates to resemble the transition state.'' For example, phosphonate **1** can be considered a mimic of the transition state for the hydrolysis of ester **2.** Indeed, antibodies obtained from immunization with TSA **1** catalyze the hydrolysis of ester **2** to give the acid (**3**) (Scheme 1) (Janda *et al.* 1989, Pollack *et al.* 1989, Tramontano *et al.* 1988).

2. Reactive Immunization

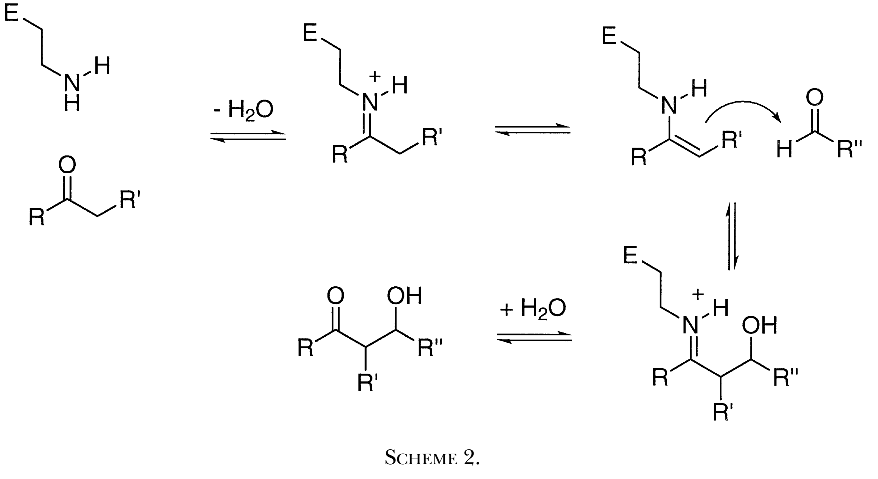
Although the transition state analog approach is suitable for enzymes that bind their transition state noncovalently, many natural enzymes achieve rate accelerations through covalent catalysis. For example, in the mechanism of most esterases and amidases, a functional group (e.g., a serine hydroxyl) of the protein covalently interacts with the substrate to form a protein bound intermediate. Furthermore, nature's most fundamental carbon–carbon bond-forming enzymes, class I aldolases, use



Scheme 1.

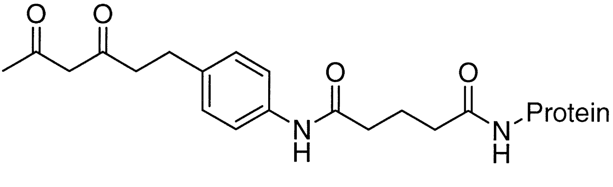
covalent catalysis as well. In these enzymes, a lysine ε-amino group interacts with ketone substrates to form Schiff bases and enamines (Scheme 2) (Lai *et al.* 1974, Morris *et al.* 1994). Given the ubiquity of the aldol reaction in both natural biosynthesis and organic synthesis, we were interested in developing antibody catalysts for the aldol reaction. Therefore, we set out to develop antibody catalysts that mimic the covalent mechanism of natural class I aldolases.

To address this sort of reaction we needed to develop a new approach to catalytic antibodies or enzymes in general that differed significantly from the noncovalent approach of Pauling and Jencks. We sought a method that would allow for the programming of detailed aspects of chemical reaction mechanisms, down to the level of the chemical identity



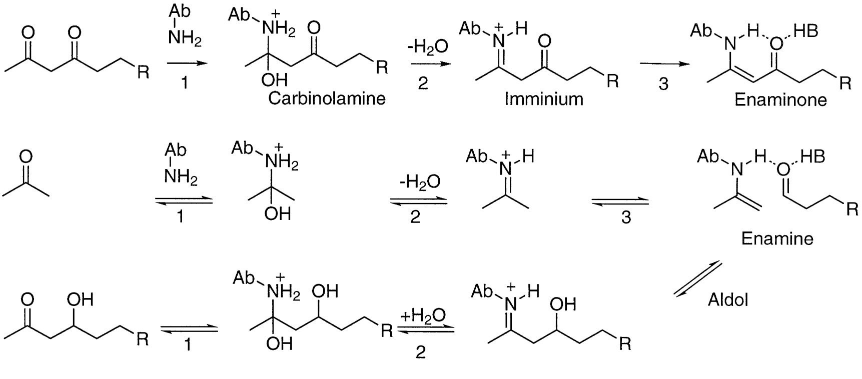
Scheme 2.

of a residue to be used in the catalytic reaction. Our proposed solution to this problem was to use an antigen that would be reactive enough to form a covalent stable conjugate with an antibody that possessed a reactive amino group. The hapten we chose was 1, 3-diketone **4** (Wagner *et al.,* 1995; Barbas *et al.,* 1997).

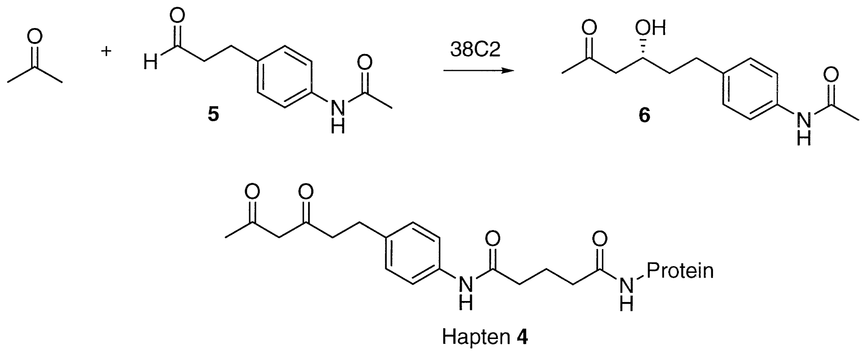


4

We anticipated that if **4** encountered a reactive amino group in the correct chemical microenvironment, it would form an enamine. In contrast to regular enamines, this enamine (or enaminone or vinylogous amide) would be a stable species because of conjugation and hydrogen bonding. Thus, unlike a transition state analog, this hapten would be covalently complementary to a low pK_a amine. The formation of the enaminone (as monitored by UV spectroscopy) would indicate that parts of the catalytic machinery, beyond the reactive amine, that were necessary for aldol catalysis are present (Scheme 3). These would include (1) catalysis of carbinolamine formation, (2) dehydration, and (3) deprotonation of the imminium intermediate. The carbon–carbon bond-forming step may further be catalyzed by an acid that protonates the second carbonyl group in the hapten:



Scheme 3.



SCHEME 4.

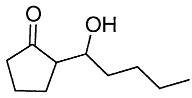
The symmetry of the aldol mechanism would then allow for catalysis of the hydrolysis of the resulting hydroxyimminium compound to the aldol.

Diketone **4** was used for immunization, and two out of twenty monoclonal antibodies produced showed the characteristic enaminone absorbtion at 315 nm after incubation with the diketone. Only these two antibodies, 38C2 and 33F12, were aldol catalysts (Wagner *et al.* 1995). For example, 38C2 catalyzes the aldol reaction between acetone and aldehyde **5** to give aldol **6** with $k_{cat} = 6.7 \times 10^{-3}$ min^{-1} and $K_M = 17$ μM (Scheme 4).

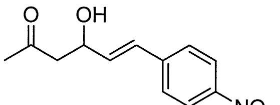
B. Aldolase Antibody 38C2

1. Scope

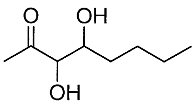
It soon became clear that aldolase antibody 38C2 catalyzed the aldol reaction with both efficiency and broad scope. For example, compounds **7** to **13** could all be synthesized in antibody 38C2 catalyzed aldol reactions (Hoffmann *et al.*, 1998).



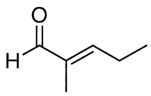
7



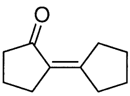
8



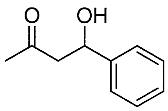
9



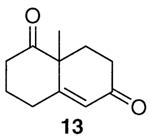
10



11



12



13

These results were unprecedented in the catalytic antibody area. Traditionally, antibody catalysts are very specific and only catalyze the reaction of a single substrate (or substrate combination). This is because in the normal case of immunization, a series of somatic mutations usually leads to an increase in binding specificity toward the inducing antigen. Antibody catalysts generated from this process have the restricted substrate specificity characteristic of most natural enzymes.

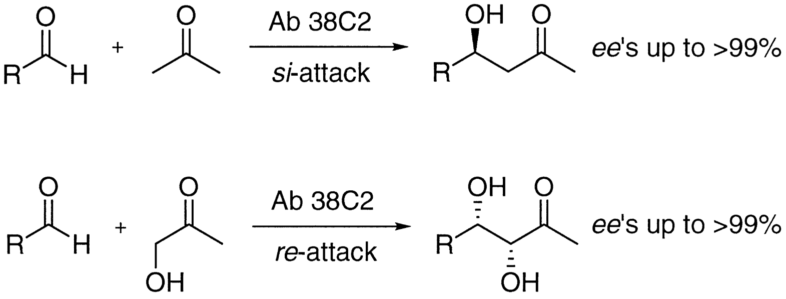
In the covalent case, the usual process of somatic refinement may be aborted because any clone that carries an antibody that has made a covalent bond with the antigen will be selected above clones containing only noncovalent antibody-antigen interactions. This is because no matter how many productive, noncovalent interactions are generated by competing clones, they cannot equal the binding energy achieved by a single covalent event. When the covalent event appears early in the process of antibody evolution, any selective pressure on the refinement process may cease. Thus, antibodies selected in this way can still be efficient because, like enzymes, they were selected on the basis of a chemical reaction. However, unlike enzymes, they are broad in scope because the usual requirement for refinement of the binding pocket has been circumvented.

2. *Enantioselectivity*

In addition to broad-scope substrate specificity, 38C2 exhibits high enantioselectivity for the aldol reaction. Although this high degree of enantioselectivity has been observed for antibody-catalyzed ester hydrolysis reactions, it is certainly not a feature common to all such catalysts (Janda *et al.,* 1989; Lo *et al.,* 1997; Pollack *et al.,* 1989; Tanaka *et al.,* 1996; Wade and Scanlan, 1996). Furthermore, the rules for the enantioselectivity for 38C2-catalyzed aldol reactions are both simple and general (Hoffmann *et al.,* 1998). For most ketone donors, attack occurs on the *si* side of the acceptor. However, when a ketone with an α-hydroxy substituent (such as hydroxyacetone) acts as donor, attack occurs on the *re* side (Scheme 5).

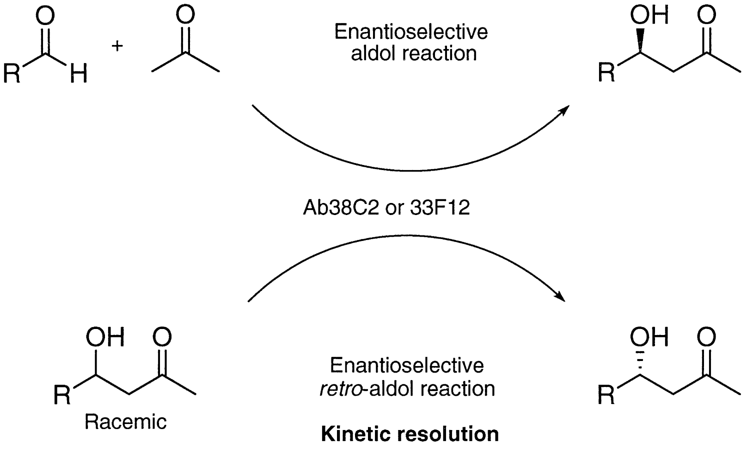
3. *Retro-Aldol Reaction*

Because an equilibrium constant is not affected by catalysis, an enzyme that accelerates a forward reaction must also accelerate the reverse or retro-reaction. Furthermore, the enantioselectivity for both reactions will be identical. Antibody 38C2 catalyzes both the forward and retro-aldol reaction, and we envisioned that it may be useful in the kinetic resolution of aldols. Because the product enantiomer from the forward aldol reaction is the substrate in the retro-aldol reaction, the opposite



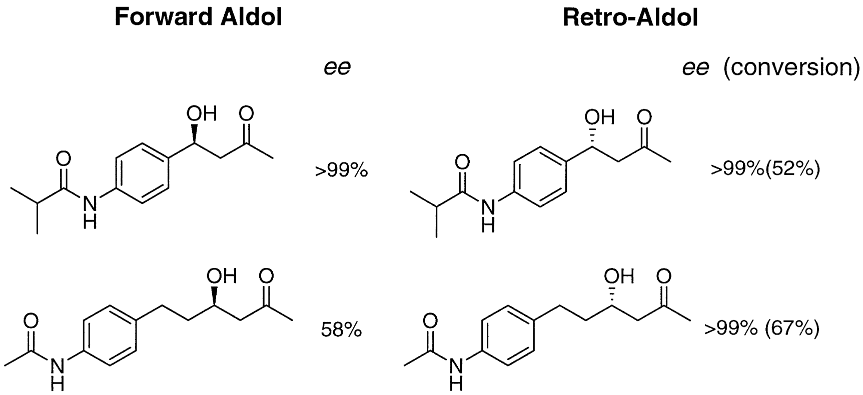
SCHEME 5.

enantiomer can be obtained in a kinetic resolution. Therefore, a single antibody catalyst could then be used for the preparation of both aldol enantiomers (Scheme 6) (Zhong *et al.,* 1998).



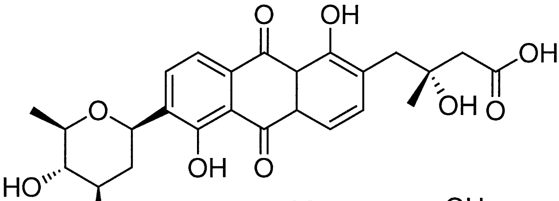
SCHEME 6.

Two examples of this are shown below (Hoffmann *et al.,* 1998; Zhong *et al.,* 1998). In the first case, both the forward and retro-aldol reactions furnished the product in high optical purity. In the second case, the forward aldol reaction gave the product with only modest *ee* (58%). However, the retro-aldol reaction provided its corresponding enantiomer in >99% *ee* after 67% conversion. These experiments demonstrate the power of kinetic resolution to provide high *ee* values in cases in which the forward aldol reaction provides only moderate *ee* values.

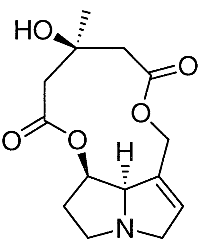


4. Tertiary Aldols

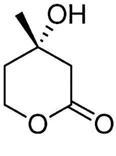
Until recently, chemical and enzymatic approaches toward the synthesis of enantiomerically enriched aldols have been applied almost exclusively to the synthesis of *secondary* aldols. Conversely, we found no general methods, either chemical or enzymic, for the preparation of enantiomerically enriched tertiary aldols, even though tertiary aldols are structural motifs common to many bioactive natural products, such as Vineomycinone B2 (**14**), Dicrotaline (**15**), Torosachrysone (**16**), Mevalonolactone (**17**), and Mycarose (**18**). Application of natural enzymes toward the synthesis of enantiomerically enriched tertiary aldols might partially be hindered by the fact that no known natural aldolase catalyzes the synthesis of tertiary aldols (List *et al.,* 1999).



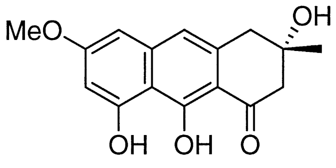
14



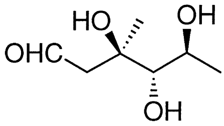
15



17



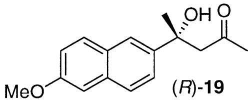
16



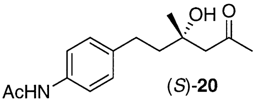
18

Although typical equilibrium constants for formation of a tertiary aldol prohibit direct forward aldol synthesis (0.002 M^{-1} for the aldol reaction of acetone with acetophenone) (Guthrie and Wang, 1992), the retro-aldol reaction is greatly favored (a 1 mM solution of the resulting tertiary aldol is converted almost completely to acetone and acetophenone at equilibrium) (List *et al.,* 1999).

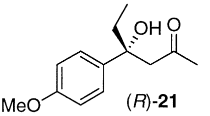
Indeed, we have found antibody 38C2 to be an efficient catalyst for the retro-aldol reaction of tertiary aldols. Aldols (*R*)−**19** through (*S*)−**26** were synthesized via kinetic resolution with 38C2 (List *et al.,* 1999). The resolution of aldols (*S*)−**24,** (*R*)−**25,** and (*S*)−**26** demonstrates the potential of aldehyde aldols. Aldehyde aldols provide facile access to acetate aldols that are otherwise difficult to obtain by more traditional techniques (Saito *et al.,* 1999).



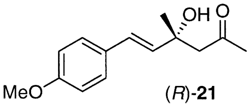
(*R*)-**19**



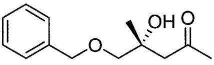
(*S*)-**20**



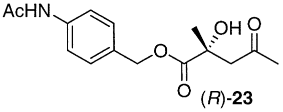
(*R*)-**21**



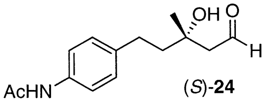
(*R*)-**21**



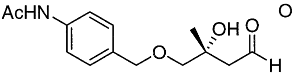
(*R*)-**22**



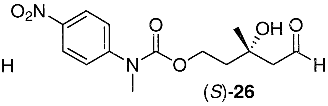
(*R*)-**23**



(*S*)-**24**



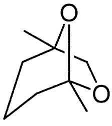
(*R*)-**25**



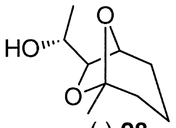
(*S*)-**26**

5. Synthetic Applications

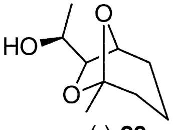
The availability of 38C2 as a broad scope, enantioselective, efficient aldolase enzyme has had a significant impact on organic synthesis. Some of the molecules we have synthesized with 38C2 include the natural products (+)−frontalin [(+)−**27**] (List *et al.,* 1999), some brevicomins [(−)−**28** and (−)−**29**] (List *et al.,* 1998a), epothilones A (**30**) and C (**31**) (Sinha *et al.,* 1998), and the Wieland-Miescher ketone [(*S*)−(+)−**32**] (Hoffmann *et al.,* 1998; Zhong *et al.,* 1997). The brevicomin examples represent the first use of a catalytic antibody to decrease the total number of synthetic steps and increase the enantioselectivity of natural product syntheses.



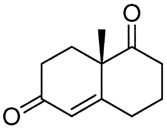
(+)-**27**



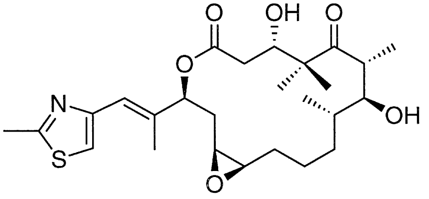
(-)-**28**



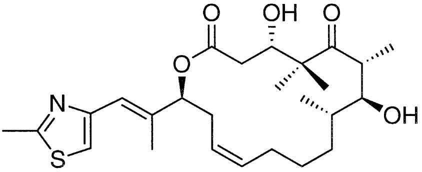
(-)-**29**



(*S*)-(+)-**32**



30



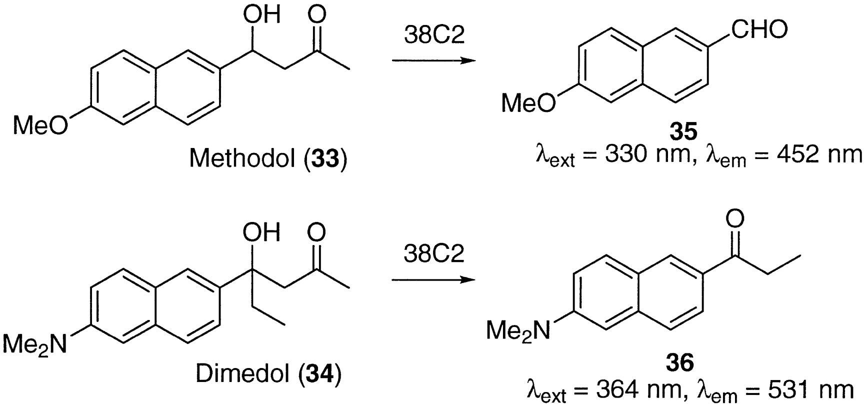
31

6. Aldol Sensors and the Tandem Retro-Aldol-Retro-Michael Reaction

Although some 38C2 catalyzed reactions have rates similar to those observed with natural class I aldolases, this is definitely not a general feature for all substrates. Indeed, some synthetically useful reactions have K_{cat}'s below 1 d^{-1}. We became interested in testing whether directed evolution would help to further improve the efficiency of our aldolase antibodies. We planned to use fluorogenic substrates (''aldol sensors'') in combination with high throughput techniques to select for catalysis in large libraries of antibodies. Although a number of sensors for hydrolytic reactions are available, there are no known systems for C–C bond-forming or cleaving reactions such as the aldol or retro-aldol reaction. We chose systems based on *retro*-aldolization that would release a fluorescent product after the reaction. Initially, we studied donor/acceptor based naphthalene systems like methodol (**33**) and dimedol (**34**) (List *et al.*, 1998b). Both sensors yield strongly fluorescent aldehydes **35** and **36** as products (Scheme 7).

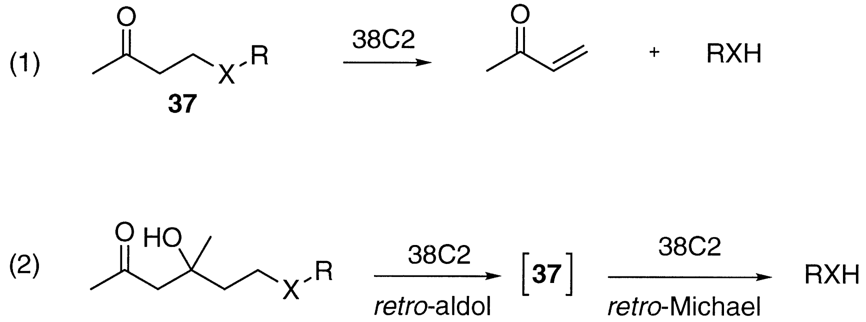
We routinely use these substrates for kinetic characterizations of new aldolase catalysts. However, they are not suitable for cell-based screenings because both substrate *and* product are readily cell permeable. The solution to this problem came when we discovered that our aldolase antibodies catalyze the *β*-elimination (or *retro*-Michael reactions) of *β*-hetero substituted ketones **37** [Scheme 8 (1)].

These reactions are generally fast and formally allow us to use any known fluorogenic, chromogenic, or luminogenic substrate, when the detectable property is based on an ionizable XH group (*p*-nitrophenol, umbelliferone, etc.) (Klein and Reymond, 1998). The antibody catalyzed

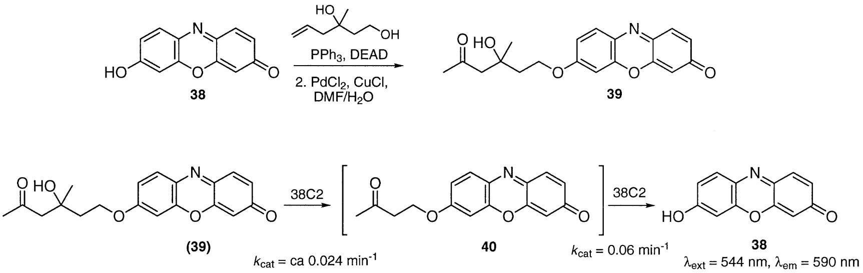

OH O
38C2
CHO
MeO
Methodol (33)
35
λ_{ext} = 330 nm, λ_{em} = 452 nm
Me2N
Dimedol (34)
36
λ_{ext} = 364 nm, λ_{em} = 531 nm

SCHEME 7.

retro-Michael reaction has a relatively high background. Because of the low background of the aldol reaction, we reasoned that a tandem retro-aldol-retro-Michael reaction [Scheme 8 (2)] should give a significantly decreased background rate of fluorescence generation compared to the retro-Michael reaction alone (List *et al.,* 1998b). The actual carbon–carbon bond-cleaving event of the retro-aldol reaction is "translated" into a carbon-heteroatom bond cleavage, which in turn leads to a detectable signal. To test this concept, we prepared a variety of chromogenic, fluorogenic, and luminogenic retro-aldol-retro-Michael sensors, which on treatment with 38C2 specifically and efficiently give the corresponding reporter molecules. We focused on resorufin (**38**) (Scheme 9) since its red fluorescence emission (590 nm) is red-shifted well beyond the autofluorescence exhibited by most biological samples. This feature,



SCHEME 8.


38
PPh3, DEAD
2. PdCl2, CuCl, DMF/H2O
39
(39)
38C2
k_{cat} = ca 0.024 min^{-1}
40
38C2
k_{cat} = 0.06 min^{-1}
38
λ_{ext} = 544 nm, λ_{em} = 590 nm

SCHEME 9.

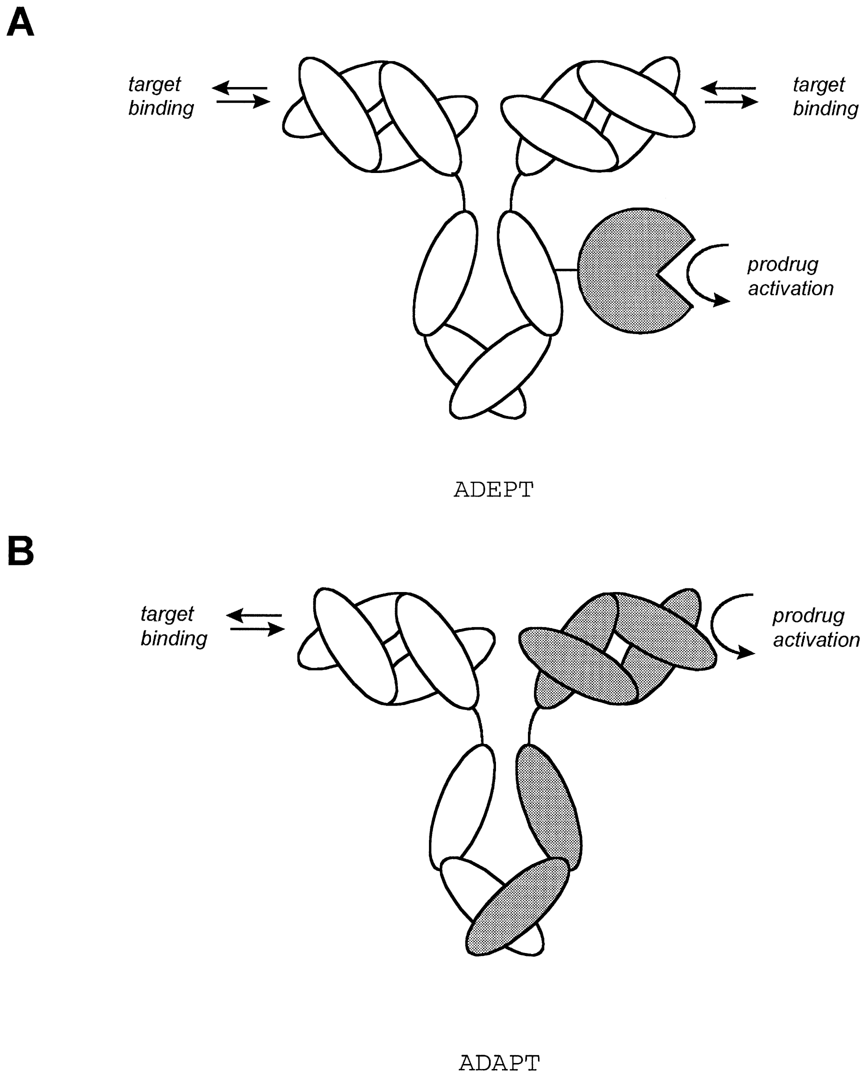
along with its nontoxicity, allows for its use in flow cytometry and other assays within living cells (Behrens *et al.,* 1998; Donato *et al.,* 1993). Retro-aldol-retro-Michael sensor **39** has been synthesized and shown to be an efficient substrate for 38C2 (List *et al.,* 1998b), although the rate at which it is processed is lower than aldols **33** and **34.** The K_{cat} of the retro-aldol-retro-Michael-tandem reaction has been estimated to be around 0.024 min^{-1} (K_M = 70 μM). However, the transformation is very specific and $k_{cat}/k_{uncat} > 10^5$. The k_{cat} for the retro-Michael reaction of ketone **40** alone was determined to be 0.06 min^{-1} (K_M = 13 μM). Here, the background reaction in PBS buffer (k_{uncat} = 0.0009 min^{-1}) is only sixty-seven times slower. This observation applies for all retro-aldol-retro-Michael sensors we have prepared to date. Currently, we use fluorogenic substrates that are based on chloromethylated fluorescein that possess both excellent cell retainment and the characteristic fluorescence of fluorescein.

7. *Prodrug Activation*

The retro-aldol-retro-Michael chemistry that we have developed in the context of fluorogenic substrates turned out to be of broad generality. In principle, any ionizable heteroatom-containing molecule may be functionalized to become a substrate for 38C2-catalyzed reaction.

Based on this reaction sequence and the ADEPT concept (antibody directed enzyme prodrug therapy) (Niculesco-Duvaz and Springer, 1997), we have developed a novel and broadly applicable drug-masking chemistry that operates in conjunction with a unique broad scope catalytic antibody. Whereas the ADEPT strategy conjugates an enzyme to an antibody (Scheme 10A), the ADAPT strategy (antibody directed abzyme prodrug therapy) uses a bifunctional antibody (Cao and Suresh, 1998; Rader and List, 2000) for both binding and catalysis (Scheme 10B).

The ADAPT concept consists of a humanized bispecific antibody with one catalytic and one tumor specific arm, in combination with a nontoxic

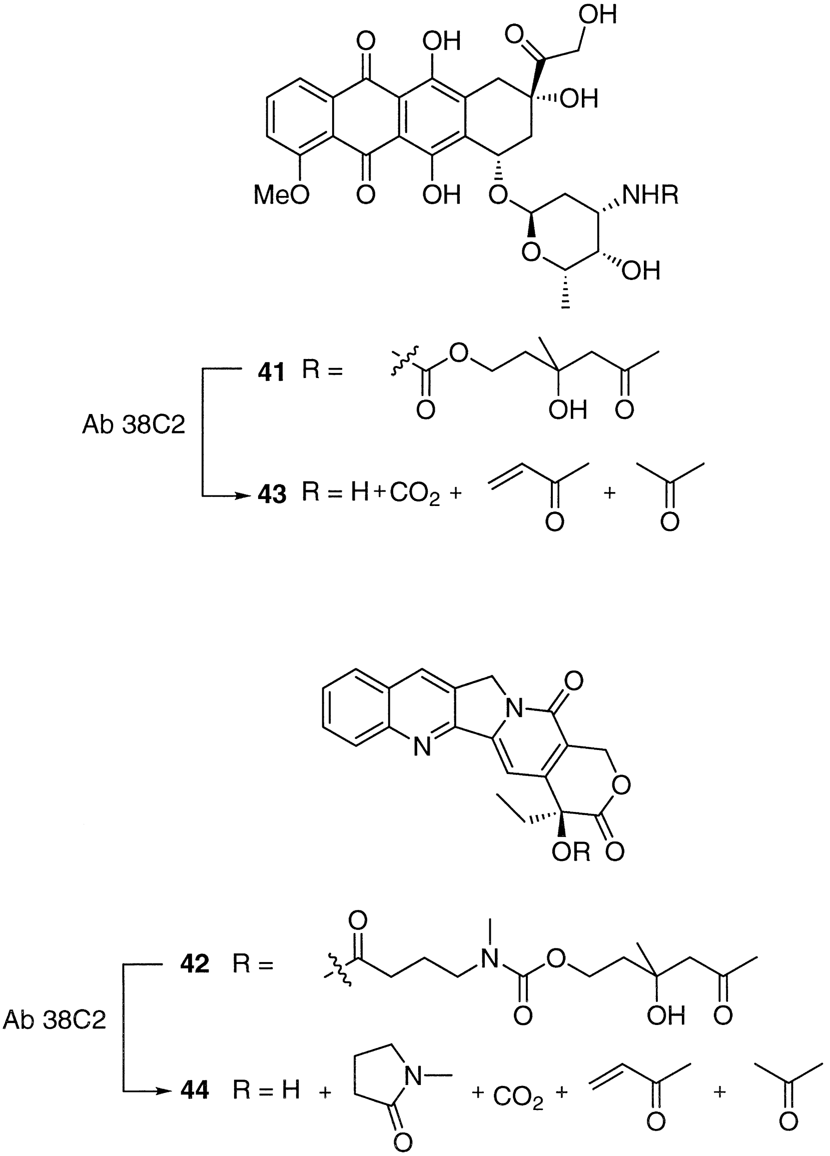


ADEPT

ADAPT

Scheme 10.

prodrug. First, the antibody is administered and allowed to bind its tumor targets. After excess antibody has been cleared from blood and periphery, a nontoxic compound is administered that possesses the retro-aldol-retro-Michael subunit. Sequential retro-aldol-retro-Michael reactions catalyzed by the catalytic part of the antibody selectively convert



Scheme 11.

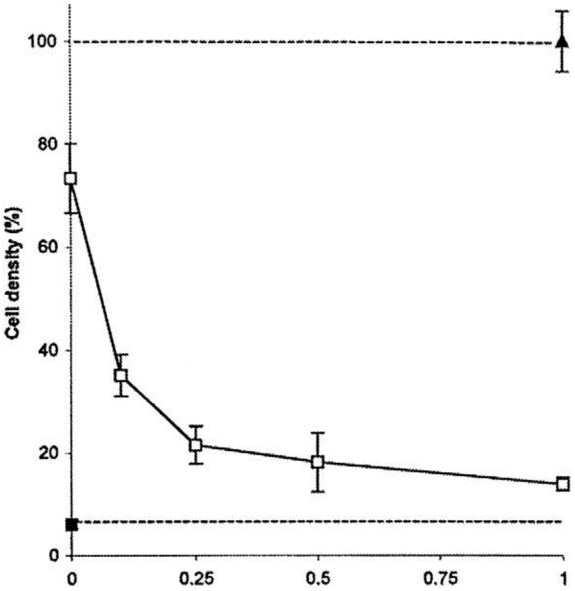


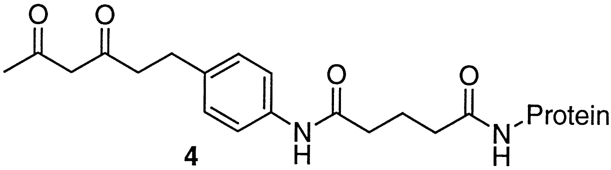
FIG. 6. Growth inhibition of HT29 human colona carcinoma cells by procamptothecin in the presence of increasing concentrations of antibody 38C2. Filled triangle: untreated control; open square: 1 μM procamptothecin; filled square: 1 μM camptothecin. Bars indicate SD, $n = 4$.

this prodrug into a cytotoxic agent. This reaction cascade is not catalyzed by any known natural enzyme. Application of this masking chemistry to the anticancer drugs doxorubicin (**43**) and camptothecin (**44**) produced prodrugs (**41** and **42**) (Scheme 11) with substantially reduced toxicity (Shabat *et al.,* 1999). The catalytic antibody selectively unmasks these prodrugs when it is applied at therapeutically relevant concentrations (Fig. 6).

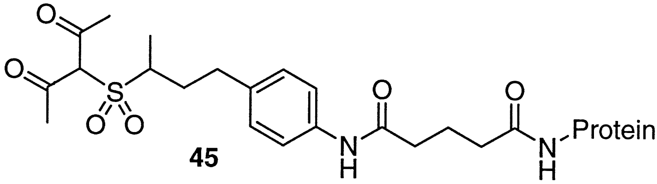
C. 84G3: Combining TSA and Reactive Immunization

1. Hapten Design

In our original hapten design for aldolase antibodies, the β-diketone functionality of hapten **4** was used as a reactive immunogen to trap a chemically reactive lysine residue in the active site of an antibody as a stable enaminone. The chemical mechanism leading up to the stabilized enaminone should match that of Class I aldolases over this portion of the reaction coordinate.

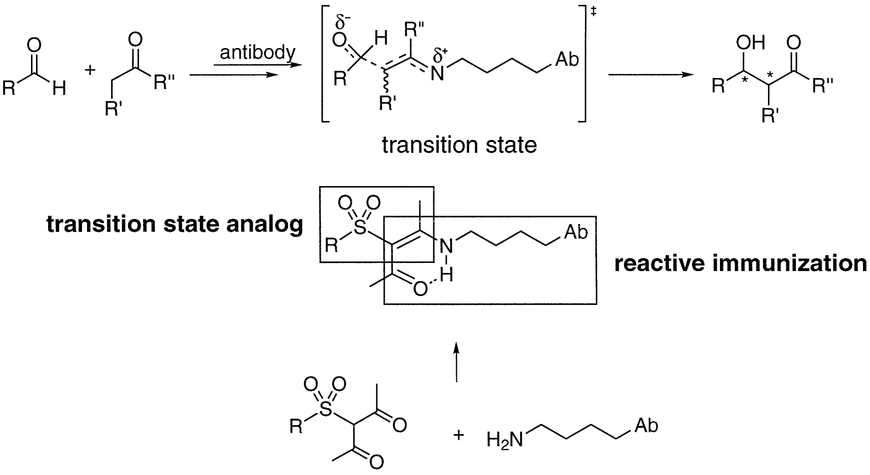


4



45

A perceived limitation of our original hapten design (**4**) is that it does not address the tetrahedral geometry of the rate-determining transition state of the C–C bond-forming step. Hapten **45** addresses this limitation and contains features common to the transition state analog approach and reactive immunization strategy. The tetrahedral geometry of the sulfone moiety in hapten **45** mimics the tetrahedral transition state of the C–C bond-forming step and therefore should facilitate nucleophilic attack of the enaminone intermediate on the acceptor aldehyde (Scheme 12). Using this new hapten, seventeen monoclonal antibodies were prepared, and nine were shown to be catalytic (Zhong *et al.,* 1999). This



Scheme 12.

ratio of catalytic to noncatalytic antibodies (9 : 17) is much greater than with hapten **4,** where only two of twenty antibodies were catalysts (Wagner *et al.,* 1995).

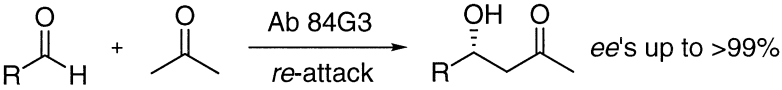
2. *Antipodal Reactivity*

The enantioselectivity of transition metal catalyzed aldol reactions is readily reversed by exchange of the chiral ligand that directs the stereochemical course of the reaction. Since a general approach to the reversal of enantioselectivity is not available with enzymes, we were pleased to discover that the enantioselectivity for 84G3-catalyzed aldol reactions was antipodal to 38C2. Whereas 38C2 catalyzes the *si*-attack of donor ketones on aldehyde acceptors, 84G3 catalyzes the *re*-attack (Scheme 13). As with 38C2, *ee*'s up to >99% are achievable with 84G3 (Zhong *et al.,* 1999).

3. *Gram-scale Syntheses*

The enantioselective aldol syntheses described previously highlight the synthetic potential of our broad-scope aldolase antibodies. However, when using catalytic antibodies for chemical synthesis, a dilemma we often encounter is that many of the interesting and synthetically useful organic molecules have low water solubility, while catalytic antibodies, like all proteins, function ideally in aqueous environments. The potential benefits of using catalytic antibodies for organic synthesis have motivated us in developing a biphasic aqueous/organic solvent system for the gram-scale retro-aldol kinetic resolution of racemic aldols using aldolase antibodies 38C2 and 84G3 (Turner *et al.,* 2000 in press).

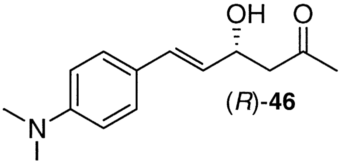
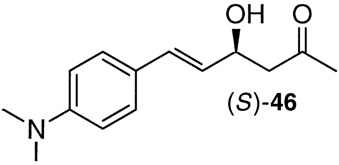
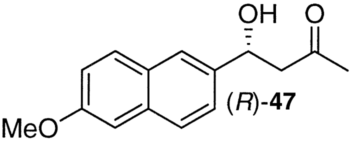
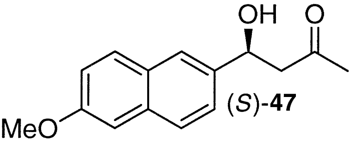
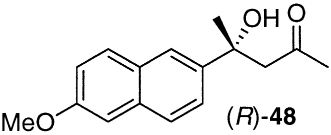
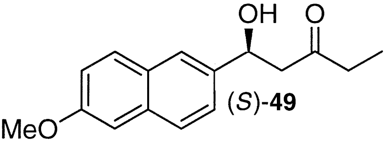
In a typical reaction, a solution of the antibody in phosphate buffered saline (PBS) is added to a solution of the racemic substrate (ca. 50–100 mM) in either toluene or chlorobenzene. The mixture is shaken, while the substrate *ee* is monitored by chiral HPLC. When the desired *ee* is reached, the reaction mixture is cooled (−20°C), allowing easy separation of the organic layer from the frozen aqueous antibody solution. The aldol product is purified by column chromatography, and the antibody solution is thawed for reuse.



SCHEME 13.

The results for our kinetic resolutions are shown in Table III. The biphasic procedure was used to synthesize enantiomerically pure aldols by kinetic resolution on a scale of miligrams to grams, with amounts of antibody binding sites ranging from 0.0086 to 0.12 mol %. To illustrate the potential for catalyst recycling, we chose to synthesize (*S*)-**47** via three sequential resolutions of aldol *rac*-**47** with the same 84G3 catalyst.

TABLE III
Enantiomerically Pure Aldols Synthesized by Antibody Catalyzed Kinetic Resolution.

Product	Antibody	Time (h)	Recovery[a]
 (*R*)-**46**	38C2 (255 mg; 0.025 mol %)	88	1.55 g (49%) >97% *ee*
 (*S*)-**46**	84G3 (16 mg; 0.015 mol %)	340	154 mg (48%) 95% *ee*
 (*R*)-**47**	38C2 (15.4 mg; 0.10 mol %)	144	25 mg (50%) 97% *ee*
 (*S*)-**47**	84G3 (210 mg; 0.065-0.068 mol %)	91 172 259	469 mg (47%) 441 mg (42%) 458 mg (43%) 97% *ee*
 (*R*)-**48**	38C2 (18 mg; 0.12 mol %)	193	22 mg (44%) 99% *ee*
 (*S*)-**49**	84G3 (500 mg; 0.0086 mol %)	65	10 g (50%) >99% *ee*

[a] Theoretical maximum for 50% recovery is 100% *ee*.

Racemic aldols **46–48** have k_{cat}'s of 1 to 5 min^{-1} (List *et al.,* 1998b) and were used for kinetic resolutions on scales ranging from mg to a few g, using low amounts of catalyst (<0.15 mol %). We believe that these rates represent the minimum substrate turnover required to be useful for preparative scale syntheses, and substrates with higher turnovers can be conveniently synthesized in even larger quantities. Indeed, aldol *rac*-**49** has a k_{cat} of 46.8 min^{-1} (Zhong *et al.,* 1999) and was resolved on a 20-g scale in a reaction volume of 700 ml. A 50% recovery and >99% *ee* was achieved using 500 mg (0.0086 mol %) of 84G3 (from our initial antibody supply of 100 g) as catalyst (this antibody supply was produced entirely from cell culture). This is the largest reaction scale ever performed with an antibody catalyst (Fig. 7). The issue of scale in antibody catalyzed reactions can be addressed with currently available technologies, particularly the heterologous expression of antibodies in plants and algae, where low-cost production of these catalysts on a multiton scale could be achieved to allow for the ''green synthesis'' of aldols on a virtually unlimited scale. Antibody 38C2 is commercially available from the Aldrich Chemical Company.

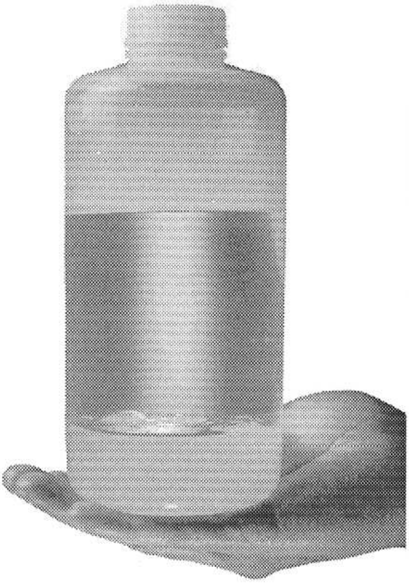
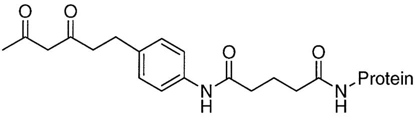


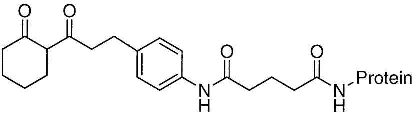
FIG. 7. Antibody catalyzed resolution of *rec*-**49** on a 20-g scale.

IV. Multiple Hapten Selection: Refining the Active Site of a Catalytic Antibody by *In Vitro* Selection

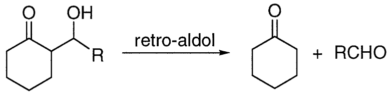
Although aldolase antibodies generated by reactive immunization are broad in scope, the rates of different substrates can vary by several orders of magnitude. In cases in which turnover rates are low for specific types of substrates, we are developing methods to increase rates by screening phage libraries for binding to multiple haptens. For example, rates for the retro-aldol reaction of **50** catalyzed by antibodies selected from phage libraries screened against both haptens **4** and **51** can be almost an order of magnitude faster than antibodies screened against hapten **4** alone. Furthermore, selection for the second hapten does not cause the antibody to ''forget'' the substrate specificity implanted by the first hapten. Selection against multiple haptens, along with the technique of reactive immunization, provide general methods to broaden the scope of catalytic antibodies.



4



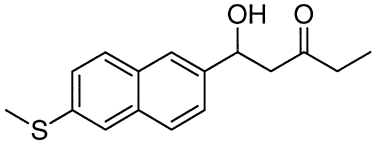
51



50

A. *The Challenge: Can Catalytic Antibodies Be Faster than Their Natural Counterparts?*

During the evolution of a natural enzyme, selection is not solely dependent on rate improvement. Therefore, there is no requirement for enzymes to be kinetically ''perfect,'' and it should be possible to develop catalytic antibodies that are faster than their natural counterparts. The designed substrate **52** has a rate of 1.4 s^{-1} with 84G3-catalyzed retro-aldol reaction (Zhong *et al.*, 1999). Its kinetic parameters hold the current world record for antibody catalysis ($K_M = 4.2\ \mu M$, $k_{cat}/k_{uncat} = 2 \times 10^8$, $(k_{cat}/K_M)/k_{uncat} = 5 \times 10^{13}$).



52

Furthermore, the catalytic efficiency (K_{cat}/K_M) of 84G3 for this substrate, 3.3×10^5 $s^{-1}M^{-1}$, is comparable with the efficiency of natural muscle aldolase, 4.9×10^4 $s^{-1}M^{-1}$, in the retro-aldolization of its substrate fructose 1, 6-bisphosphate (Morris and Tolan, 1994). However, these two enzymes use different substrates, and the rates were recorded at different temperatures (22°C for 84G3, 4°C for the natural aldolase). Despite this, we believe that it will be possible to develop a catalytic antibody that, under identical conditions, has a faster k_{cat} and lower K_M than a natural enzyme for the same substrate.

V. Selection and Evolution of Novel DNA-Binding Proteins: From Principles to Applications

The ability to manipulate the expression of specific endogenous genes would have wide-ranging applications in medicine and experimental and applied biology. In the therapy of cancer, for example, oncogenes, angiogenic factors, metastatic factors, and drug resistance genes are all potential targets for down regulation, while tumor suppressors, immunostimulants, and apoptotic factors could be targets for specific up regulation. Down regulation of viral genes or host receptors could impact virus-related diseases such as AIDS, up regulation of fetal hemoglobin might impact sickle-cell anemia, and controlled regulation of insulin or dopamine synthesis would have therapeutic potential for diabetes and neurological disorders. A number of promising approaches for the control of gene expression have been described, operating either at the transcriptional level, such as polyamides, or the posttranscriptional level, such as antisense and ribozymes (Gottesfeld *et al.,* 1997; Matteucci and Wagner, 1996; Zaug *et al.,* 1986). We have instead decided to focus on and embellish the regulatory strategy developed by eons of evolution, that of sequence-specific transcription factors.

Transcription factors are modular proteins consisting typically of a DNA-binding domain that provides for localization of the protein to a specific DNA address, and an effector domain that directs the type of activity to take place at the site (Cowell, 1994; Ptashne, 1997). The wide range of effector activities available provide significant advantages over other proposed technologies. Transcriptional repression and activation are just two of the effector activities that are readily accessed; others include covalent modifications such as DNA methylation and cleavage. Potent activation domains, such as the herpes simplex virus VP16 (Sadowski *et al.,* 1998), and repression domains, such as the human Krüppel-associated box (KRAB) (Margolin *et al.,* 1994), have been shown to regulate a variety of promoters in a location and orientation independent fashion when fused to a sequence-specific DNA-binding protein such as the yeast GAL4. The primary challenge has therefore been to create novel sequence-specific DNA-binding proteins with sufficient specificity and affinity to target an effector domain to a unique gene among the 100,000 in the human genome. A decade of research toward this goal has finally produced the first proteins capable of such recognition, using the zinc finger DNA-binding motif (Segal and Barbas, 2000).

The His_2-Cys_2 zinc finger domain is the most common DNA binding motif found in nature (Rhodes and Klug, 1993). Over 2000 domains have been identified, and it is estimated that as much as 1% of the human genome codes for these proteins. However, the feature that makes zinc finger proteins particularly well suited for the production of gene-specific regulators is that their domains are arranged as covalent tandem repeats. A single zinc finger domain consists of approximately thirty amino acids with a simple $\beta\beta\alpha$ fold stabilized by hydrophobic interactions and the chelation of a zinc ion between to histidines and two cysteines (Fig. 8) (Lee *et al.,* 1989; Miller *et al.,* 1985). These domains

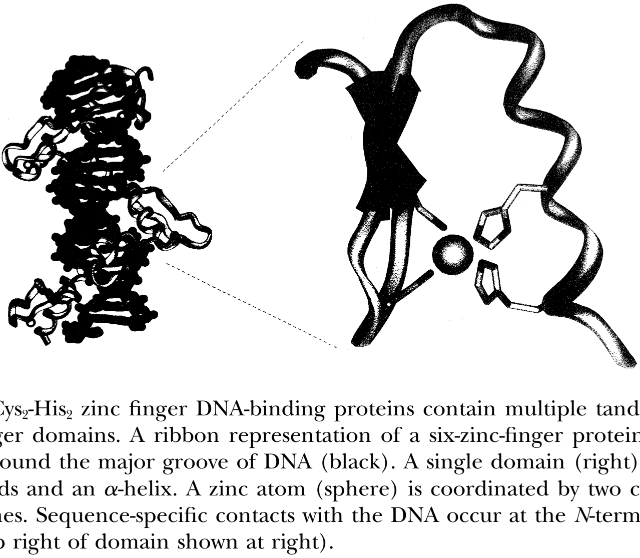


FIG. 8. Cys_2-His_2 zinc finger DNA-binding proteins contain multiple tandem repeats of zinc finger domains. A ribbon representation of a six-zinc-finger protein (white) is wrapped around the major groove of DNA (black). A single domain (right) consists of two β-strands and an α-helix. A zinc atom (sphere) is coordinated by two cystines and two histidines. Sequence-specific contacts with the DNA occur at the *N*-terminus of the α-helix (top right of domain shown at right).

can be found in arrays of up to thirty-seven repeats (Rhodes and Klug, 1993), facilitating perhaps recognition of extended DNA sequences. However, it should be noted that sequence-specific binding of more than three contiguous zinc finger domains within a naturally occurring polydactyl proteins has yet to be observed. The polydactyl nature of zinc finger proteins offers a distinct advantage over many other DNA-binding motifs that form homo- or hetero-dimers, and are therefore limited in their recognition potential.

A. *Domains of Novel Specificity*

Presentation of the zinc finger α-helix into the major groove of DNA allows for sequence-specific base contacts, with each domain typically recognizing three nucleotides. The first crystal structure of a zinc finger protein bound to DNA, that of the murine transcription factor Zif268, suggested specific roles for each residue in the recognition helix (Elrod-Erickson *et al.,* 1996; Pavletich and Pabo, 1991). With respect to the start of the α-helix, positions -1, 3, and 6 contacted the $3'$, middle, and $5'$ nucleotides, respectively. Positions -2, 1, and 5 were often involved in direct or water-mediated contacts to the phosphate backbone. Position 4 was typically a leucine, a highly conserved residue that packs in the hydrophobic core of the domain. Position 2 has been shown to interact with other helix residues and with bases depending on the helical and DNA sequences. The interactions between fingers were limited, and the interactions of each domain with the DNA were fairly restricted to one strand of a three-nucleotide site (with one notable exception, described below). Although subsequent structures of naturally occurring zinc finger proteins, such as GLI (Pavletich and Pabo, 1993) and YY1 (Houbaviy *et al.,* 1996) would reveal more complex and cooperative recognition paradigms, many proteins, such as SP1 (Narayan *et al.,* 1997), TTK (Fairall *et al.,* 1993), and TFIIIA (Nolte *et al.,* 1998; Wuttke *et al.,* 1997), were found to follow the same general pattern of Zif268.

Understanding these unique properties led us and others to experimentally manipulate the specificity zinc finger domains, using rational or combinatorial methods to modify residues in the recognition helix (reviewed in Segal *et al.,* 1999). Early hopes that the interactions could be reduced to a simple one amino acid:one nucleotide recognition code have proven unsustainable (Corbi *et al.,* 1998; Wolfe *et al.,* 1999; and Segal *et al.,* 1999, Segal and Barbas, 2000). Some of the most informative data concerning sequence-specific recognition with this class of proteins has been generated using phage display technology. Our studies have utilized a library of three-finger proteins, in which the helical residues of one finger have been randomized (Segal *et al.,* 1999; Wu *et al.,* 1995). Selection for binding is done in a way analogous to our studies using phage displayed antibodies. More specifically, residues in finger 2 of a Zif268-like protein were randomized to create a library of approximately 10 billion proteins. We then selected these proteins to recognize each one the sixteen members of the 5′-GNN-3′ set of DNA sequences using a solution binding protocol (Segal *et al.,* 1999) (Fig. 9). The oligonucleotides used for selection presented DNA-binding sequences for the two unmodified fingers, 6 bp, and a third site for the binding of the mutant protein domain, 3 bp. Specific competitor DNA was introduced to select against proteins that bound to targets differing by only one or two nucleotides of nine. After eight rounds of selection, the proteins demonstrated striking conservation of all three primary DNA contact positions (−1, 3, and 6) among virtually all the clones selected to bind a given

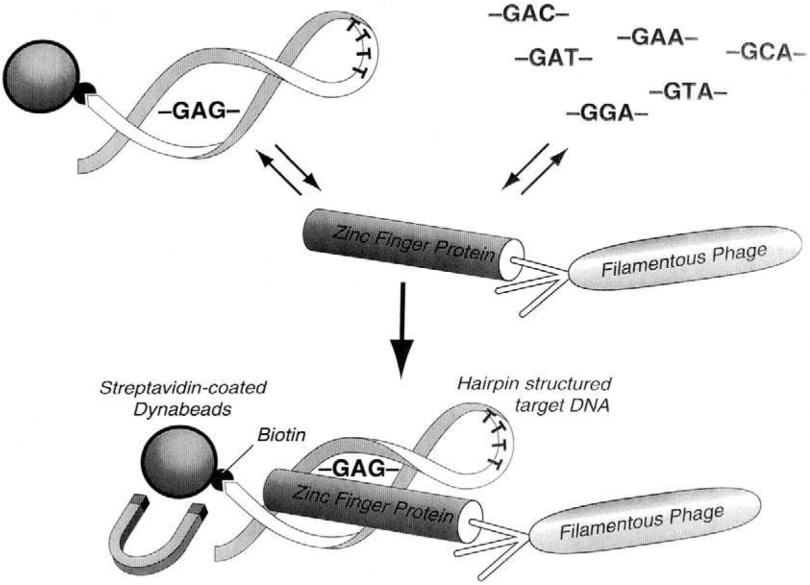
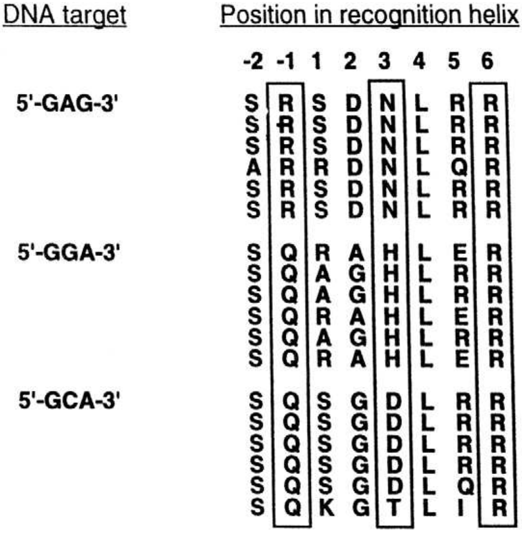


FIG. 9. Cartoon diagram of phage display and selection of zinc finger proteins. A library of three-domain zinc finger proteins is displayed on filamentous phage. In a solution binding reaction, phage are exposed to biotinylated hairpin target oligonucleotides (top left) and nonbiotinylated competitor oligonucleotides (top right), which differ by only one or two nucleotides. Phage bound to biotinylated oligonucleotides are subsequently captured on streptavidin-coated magnetic beads and immobilized with a magnet while unbound phage are washed away. Immobilized phage are then eluted, amplified, and prepared as input for the next round of selection.

target (Fig. 10). Although many of these residues were observed previously at these positions following selections with much less complete libraries, the extent of conservation we observed represented a dramatic improvement over earlier studies (Choo and Klug, 1994a; Greisman, and Pabo, 1997; Jamieson *et al.,* 1994; Jamieson *et al.,* 1996; Rebar and Pabo, 1994; Wu *et al.,* 1995). Typically, phage selections have shown a consensus selection in only one or two of these positions. The greatest sequence variation occurred at residues in positions 1 and 5, which do not make base contacts in the Zif268/DNA structure and were expected not to contribute significantly to recognition (Elrod-Erickson *et al.,* 1996; Pavletich and Pabo, 1991). Variation in positions 1 and 5 also implied that the conservation in the other positions was due to their interaction with the DNA and not simply the fortuitous amplification of a single clone for other reasons.

Impressive amino-acid conservation was observed for recognition of the same nucleotide in different targets and the selected residues in many cases made good chemical sense (Fig. 10). For example, Asn in position 3 (Asn3) was virtually always selected to recognize adenine in the middle position, whether in the context of GAG, GAA, GAT, or GAC. Gln-1 and Arg-1 were always selected to recognize adenine or



DNA target	Position in recognition helix							
	-2	-1	1	2	3	4	5	6
5'-GAG-3'	S	R	S	D	N	L	R	R
	S	R	S	D	N	L	R	R
	S	R	S	D	N	L	R	R
	A	R	R	D	N	L	Q	R
	S	R	S	D	N	L	R	R
	S	R	S	D	N	L	R	R
5'-GGA-3'	S	Q	R	A	H	L	E	R
	S	Q	A	G	H	L	R	R
	S	Q	A	G	H	L	R	R
	S	Q	R	A	H	L	E	R
	S	Q	A	G	H	L	R	R
	S	Q	R	A	H	L	E	R
5'-GCA-3'	S	Q	S	G	D	L	R	R
	S	Q	S	G	D	L	R	R
	S	Q	S	G	D	L	R	R
	S	Q	S	G	D	L	R	R
	S	Q	S	G	D	L	Q	R
	S	Q	K	G	T	L	I	R

FIG. 10. Results of phage selections of zinc finger domains for different DNA targets. The sequence of the finger-2 recognition helix for six randomly chosen clones (right) that were selected for binding to the indicated DNA target (left). Amino acid positions are in reference to the start of the α-helix. Putative contact positions −1, 3, and 6 are boxed.

guanine, respectively, in the 3′ position regardless of context. Amide side chain based recognition of adenine by Gln or Asn is well documented in structural studies as is the Arg guanidinium side chain to guanine contact with a 3′ or 5′ guanine (Elrod-Erickson *et al.,* 1996; Elrod-Erickson *et al.,* 1998; Fairall *et al.,* 1993; Kim and Berg, 1996). Often, however, two or three amino acids were selected for nucleotide recognition. His-3 or Lys-3 (and to a lesser extent, Gly-3), for example, were selected for the recognition of a middle guanine.

Rigorous study of purified proteins for their ability to recognize each of the 16 5′-GNN-3′ finger-2 subsites using a multitarget ELISA assay proved to be essential in understanding the nature of the selected proteins (Fig. 11). This assay provided an extremely sensitive test for specificity since there were always six ‘‘nonspecific’’ sites that differed from the ‘‘specific’’ site by only a single nucleotide out of a nine-nucleotide target. Many of the phage-selected finger-2 proteins showed exquisite specificity, while others demonstrated varying degrees of cross-reactivity. Indeed some proteins actually bound better to subsites other than those for which they were selected. These results underscore the common misconception that any sequence strongly selected during phage display must be the optimal sequence. The flaw in that logic is that the researcher is not always aware of what the actual selection pressure is during an experiment. A rigorous investigation of the binding specificity of the

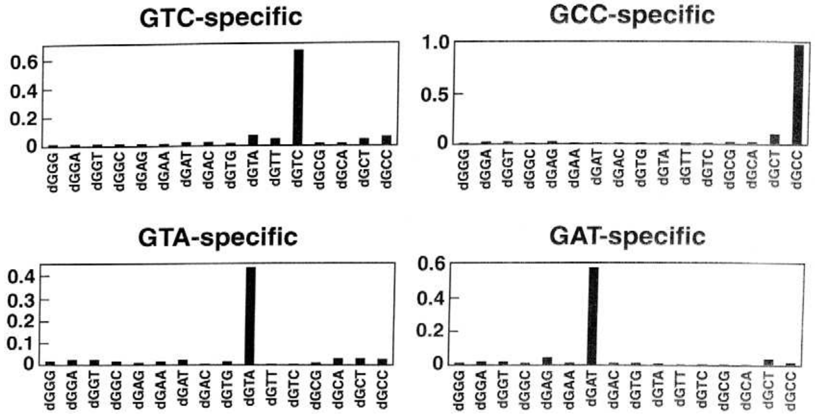


FIG. 11. Exquisite sequence specificity of zinc finger domains. The multitarget ELISA titration assay for binding specificity is described in Segal *et al.,* 1999. Black bars represent target oligos with different finger-2 subsites. The height of each bar represents the relative affinity of the protein for each target, averaged over two independent experiments and normalized to the highest signal.

new three-finger proteins revealed that some residues were selected during panning because, for example, they increased the affinity of the interaction at the cost of specificity. Indeed, virtually all studies of zinc finger domains selected by phage display have reported some domains that were selected to recognize one target but, on analysis, actually preferred binding to a different target.

We therefore found it necessary to optimize the output of phage display using site-directed mutagenesis. Nearly 100 systematic modifications of the phage-selected helices were analyzed. This data showed that many amino-acid combinations or conservative substitutions had dramatic and unexpected effects on specificity. For example, while helix positions 1 and 5 are not expected to play a direct role in DNA recognition, the best improvements in specificity always involved modifications in these positions. These residues have been observed in the structures of other proteins to make phosphate backbone contacts, which contribute to affinity in a non-sequence-specific manner. We suspect that removal of nonspecific contacts increases the importance of specific contacts to the overall stability of the complex, thereby enhancing specificity. It needs to be emphasized that improvements by modifications involving positions 1 and 5 could not have been predicted by existing "recognition codes" (Desjarlais and Berg, 1992; Suzuki *et al.,* 1994; Choo and Klug, 1994b; Choo and Klug, 1997), which typically only consider positions −1, 2, 3, and 6. Further, in contrast to the expectations of a simple recognition code, we determined that rather than the simple identity of the position −1 residue, sequence motifs at positions −1, 1, and 2 are required for highly specific recognition of the 3′ base. These residues may in fact provide the proper stereochemical context for interactions of the helix both in terms of recognition of specific bases and in the exclusion of other bases, with the net result being highly specific interactions. It appears that even in a structurally well-characterized system like zinc finger proteins, a combination of selection and site-directed mutagenesis studies is required to begin to fully understand the intricacies of protein/DNA recognition. Although our understanding of sequence-specific recognition using zinc finger proteins is not yet fully mature, these studies have resulted in the preparation of a collection of sixteen well-characterized domains recognizing each of the possible 5′-GNN-3′ binding sites. A number of these domains discriminate between sequences that differ by one in nine bases with >100-fold loss in affinity (Segal *et al.,* 1999).

B. *From Domains to Polydactyl Proteins*

Specific delivery of a DNA-binding protein to a single site within the 3.5 billion bp human genome requires an address of at least 16 bp.

Statistically, assuming random base distribution, a unique 16-bp sequence will occur only once in 4^{16} or 4.3 billion nucleotides, roughly the same size of a human genome (3.5×10^9 bp). An 18-bp address would therefore be specific within 68 billion base pairs of sequence. Theoretically, an 18-bp address could be specified by a zinc finger protein containing six domains. In practice, however, there were several caveats. The domains would need to be modular and independent; that is, they would need to bind the same three nucleotide sequence in the terminal finger-6 position as they did when they were selected in the finger-2 position, and they should not interfere with or be influenced by the neighboring fingers. It was also unknown if the periodicity of the protein domains would match that of the DNA over this extended sequence. Although natural proteins containing long polydactyl arrays of zinc finger domains had been inferred from sequence, zinc finger proteins had yet to demonstrate such long, contiguous interactions with DNA.

During our domain selection and modification study, it became clear that the domains were not completely independent. An early clue was the difficulty that every laboratory encountered in selecting fingers from Zif268-based libraries that could recognize sequences of the type 5′-ANN-3′ or 5′-CNN-3′. A consensus in the field emerged that when aspartate appeared in position 2 of one α-helix, it contacted the binding site of the finger next to it, forcing recognition at that neighboring site to be 5′-GNN-3′ or 5′-TNN-3′ (Elrod-Erickson *et al.,* 1996; Isalan *et al.,* 1997). This interaction, referred to as target site overlap, was particularly important because it extended the recognition subsite for Asp2-containing helices from three nucleotides to four. The inability to select domains recognizing 5′-ANN-3′ or 5′-CNN-3′ type sequences was subsequently understood to be due to the Asp2 of finger 3 in Zif268, forcing recognition of only 5′-(G/T)NN-3′ sequences in the finger-2 subsite. Of greater concern was that our Asp2-containing optimized fingers might have limited modularity, because they would require that each subsite be followed by a T or G. Asp2 proved essential for DNA binding in these domains. However, Asp2 occurs in only four of the sixteen fingers, and the use of contiguous 5′-GNN-3′ recognition domains circumvents the problem because all the subsites in the target start with G. In unpublished studies we have developed novel methods to select domains that are specific for the other trinucleotide sets, ANN, CNN, and TNN (Dreier *et al.,* unpublished).

The next concern was stitching the fingers together into polydactyl proteins. Among the many zinc finger proteins that have been characterized, the most useful scaffolds for building proteins of novel specificity have been those of the three-finger Zif268 and the structurally related Sp1 (Krizek *et al.,* 1991). Both have relatively limited interdomain cooperative interactions, and all three domains interact with the DNA in essentially the same way. This is true even for Zif268 or Sp1 proteins modified by selection or rational design. Crystal structures of these mutants with their cognate DNA reveal that reorientation of the helix relative to the DNA is sometimes required to achieve the appropriate interactions, but the roles of the amino acids are essentially unchanged, suggesting that these scaffolds were suitable for the display of modified helices (Elrod-Erickson *et al.,* 1998; Kim and Berg, 1996). We therefore grafted the residues from our 5′-GNN-3′ helices into an Sp1 framework (Zif268 produced similar results) (Fig. 12), and joined the two three-finger proteins with a consensus zinc finger linker (Fig. 13).

The resulting six-finger proteins have been shown to bind their 18-bp DNA targets with subnanomolar affinity (Beerli *et al.,* 1998; Beerli *et al.,* 2000; Liu *et al.,* 1997). Mutating either half of the target site resulted in a 100-fold loss in affinity, and mutating only 3-bp produced a thirty-

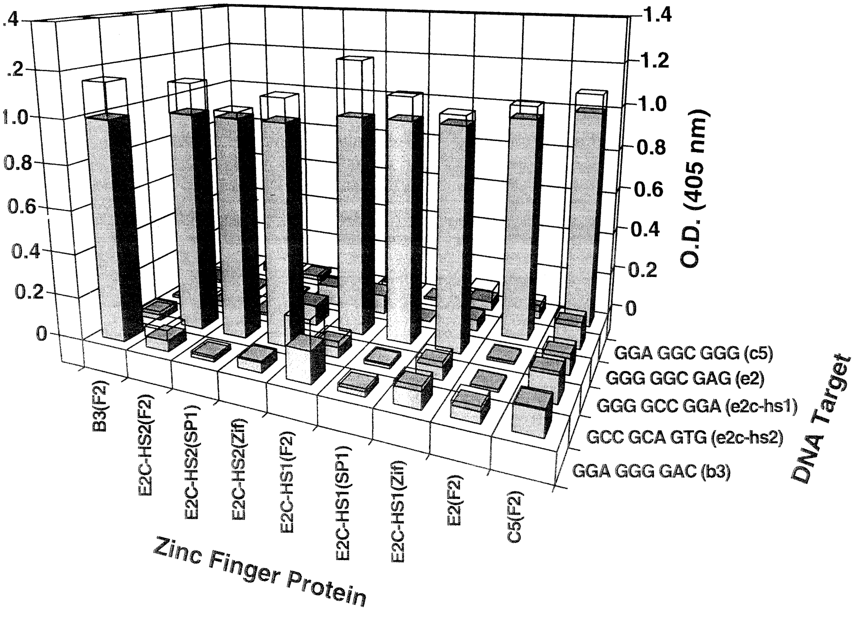


FIG. 12. Novel DNA-binding specificities of three-finger proteins. The binding of several stitched three-finger proteins constructed from predefined domains to oligonucleotides containing the indicated 9-bp targets sequences is shown. Assays were performed in duplicate and the maximal signals were normalized to 1. The open box on top of each bar represents the standard deviation. Reprinted with permission from Beerli *et al.,* (1998).

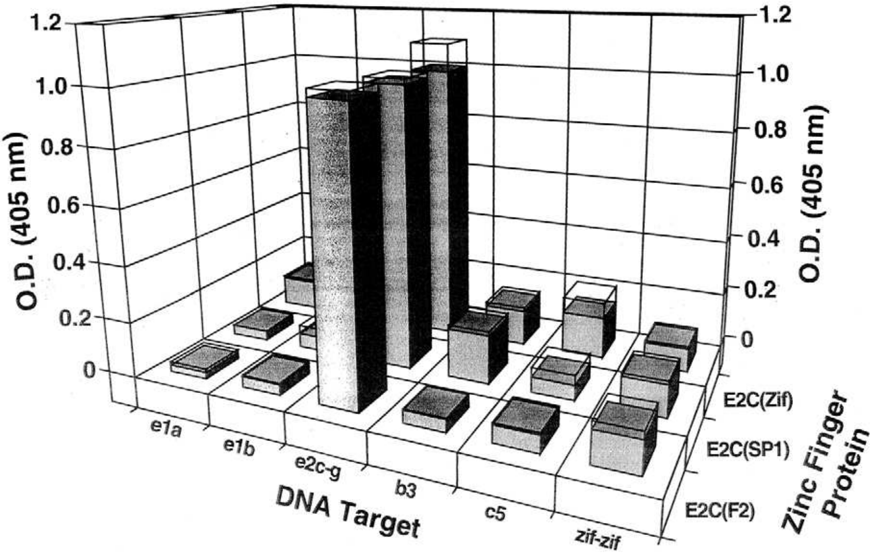


FIG. 13. DNA-binding specificities of novel six-finger proteins. Binding of a six-finger protein to oligonucleotides containing the indicated 18-bp targets was measured. The recognition helices were grafted into three different zinc-finger scaffold proteins (F2, Zif, or Sp1). The nucleotide sequences of the six-finger oligonucleotides were: e1a, 5′ -GCC GAG GCG GCC GGA GTC-3′; e1b, 5′ -GTT GTG GCG TTG GCG GCG-3′; e2c-g, 5′ -GGG GCC GGA GCC GAC GTG-3′; b3, 5′ -GCC TGA GAG GGA GCG GTG-3′; c5, 5′-GCG GAG GCA GGA GGC GGG-3′; zif-zif, 5′-GCG TGG GCG GCG TGG GCG-3′. Assays were performed in duplicate and the maximal signals were normalized to 1. The open box on top of each bar represents the standard deviation. Reprinted with permission from Beerli *et al.* (1998).

fold loss in affinity. Target site selection studies demonstrated that the specificity of these polydactyl proteins is as good or superior to those produced by other methodologies (Wolfe *et al.,* 1999; unpublished data). Moreover, for the first time, any laboratory with cloning capabilities could access the 17 million novel proteins that bind the 5′-(GNN)$_6$-3′ family of 18-bp DNA sites. Studies are ongoing to extend recognition to sequences of the type 5′-ANN-3′, 5′-TNN-3′ and 5′-CNN-3′, which would enable binding to any desired DNA sequence. However, assuming random base distribution, and considering that a unique 18-nt site can appear on either DNA strand, a 5′-(GNN)$_6$-3′ site should occur, on average, once every 2,048-bp. Therefore, the current family of 5′-(GNN)$_6$-3′ recognition proteins is capable of targeting virtually any gene in the human genome.

C. *From Polydactyl Proteins to Transcription Factors*

The human protooncogenes *erbB-2/HER-2* and *erbB-3* were chosen as model targets for the development of zinc finger-based transcriptional switches. Members of the ErbB receptor family play important roles in the development of human malignancies. In particular, *erbB-2* is overexpressed as a result of gene amplification and/or transcriptional deregulation in a high percentage of human adenocarcinomas arising at numerous sites, including breast, ovary, lung, stomach, and salivary gland (Hynes and Stern, 1994). Increased expression of ErbB-2 leads to constitutive activation of its intrinsic tyrosine kinase and has been shown to cause the transformation of cultured cells. There is increasing evidence that ErbB-3 is also involved, presumably by acting cooperatively with ErbB-2 (Alimandi *et al.,* 1995; Kraus *et al.,* 1993; Siegel *et al.,* 1999). Numerous clinical studies have shown that patients bearing tumors with elevated ErbB-2 expression levels have a poorer prognosis (Hynes and Stern, 1994). In addition to its involvement in human cancer, *erbB-2* plays important biological roles, both in the adult and during embryonal development of mammals (Altiok *et al.,* 1995; Hynes and Stern, 1994; Lee *et al.,* 1995).

Target sites of the 5′-(GNN)$_6$-3′ type were chosen in the *erbB-2* and *erbB-3* genes, designated e2c and e3, respectively (Beerli *et al.,* 1998; Beerli *et al.,* 2000). A BLAST sequence similarity search of the GenBank database confirmed that these 18-bp sequences were indeed unique, although they differed from each other by only three mismatches. Six-finger proteins constructed to recognize e2c and e3 bound their cognate targets with an affinity of 0.75 nM and 0.35 nM, respectively, and had an affinity of about 10 nM for their non-cognate targets that presented changes at three nucleotide positions. Both target sites were located in the 5′ UTR of their genes, which allowed the examination of repression through inhibition of either transcription initiation or elongation. Further, if we could demonstrate the ability to modulate gene expression by targeting within the transcribed region of a gene, EST data might be used in the design of transcriptional regulators, obviating the need to sequence the promoter region to regulate a gene.

The potential of the polydactyl proteins to function as transcriptional regulators was first tested in cell culture assays (Beerli *et al.,* 1998). HeLa cells were transiently cotransfected with zinc finger expression vectors and constructs containing a luciferase reporter gene driven by a fragment of the *erb-2* promoter. A 1.4-fold repression of luciferase activity was observed when the e2c binding protein E2C was expressed with no effector domain, suggesting that elongation by RNA polymerase was only modestly disrupted by this protein. However, expression of the zinc finger protein fused to a KRAB repressor domain (E2C-KRAB) reduced luciferase activity to undetectable (background) levels, while fusion to a modified VP16 activator domain (E2C-VP64) produced a 27-fold activation.

A greater challenge was to demonstrate regulation of endogenous genes. Endogenous genes are packaged within chromatin and are controlled by a multiplicity of *cis*- and *trans*-acting factors, raising concern as to whether specific gene regulation by a designed transcription factor would be possible. The human carcinoma cell line, A431, was transduced with a retrovirus encoding a zinc finger-effector fusion construct (Beerli *et al.,* 2000). Expression of E2C-KRAB was found to reduce endogenous ErbB-2 expression to undetectable levels, while having no effect on ErbB-3 (Fig. 14). Conversely, expression of E3-KRAB reduced endogenous ErbB-3 expression to undertectable levels, but not affecting ErbB-2. Similarly, E2C-VP64 or E3-VP64 showed upregulation (approximately 8-fold) of only their respective gene products, while having no effect on the non-target gene. These data were confirmed by flow cytometry, Western, and Northern analysis, and were also shown for other ErbB-2-expressing cell types. Using the Tet-Off™ Gene Expression System by Clontech, integrated genes for E2C-KRAB and E2C-VP64 were put under control of the tetracycline response element. Removal of the tetracycline analog, doxycycline, from the growth media resulted in expression of the zinc finger fusions and appropriate regulation of *erbB*-2. These results proved that regulation was not due to a transfection artifact, and that endogenous *erbB*-2 could be regulated by a small molecule drug. Finally, it was shown that E2C-KRAB could specifically inhibit cell-cycle progression in ErbB-2-overexpressing tumor cells, suggesting the potential of designed transcription factors for cancer gene therapy.

D. Prospects for Polydactyl Proteins

In the past few years we have demonstrated a technology for the rapid assembly of novel DNA-binding proteins that can bind a unique site in virtually any gene in the human genome with high specificity and biologically relevant affinity. When fused with appropriate effector domains, these proteins have demonstrated the capacity to up and down regulate the transcription of endogenous genes, while having no effect on other related genes. Although more effort will be needed to gain the ability to recognize any arbitrary sequence, it is important to emphasize that a readily accessible and functional technology for creating gene-specific DNA-binding proteins now exists. Future experiments will move

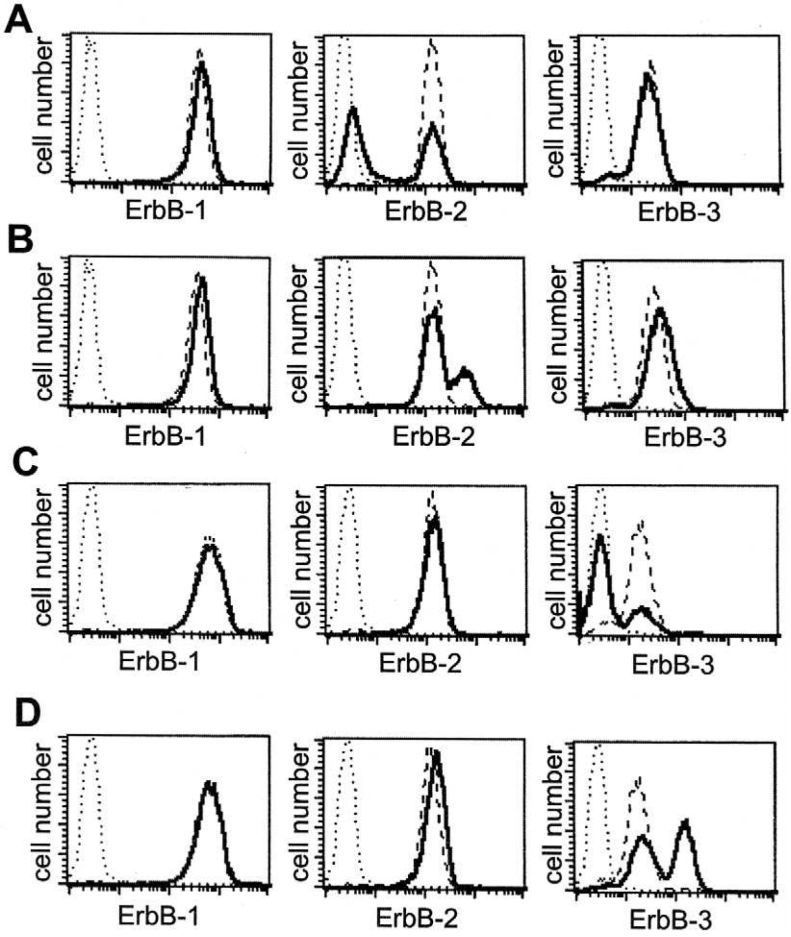


Fig. 14. Retrovirus-mediated *erbB*-2 and *erbB*-3 gene targeting. A431 cells were infected with (a) E2C-KRAB-, (b) E2C-VP64-, (c) E3-KRAB-, or (d) E3-VP64-encoding retrovirus. Three days later, intact cells were stained with an ErbB-1-specific, ErbB-2-specific, or the ErbB-3-specific mAb in combination with phycoerythrin-labeled secondary antibody, and analyzed by flow cytometry. Dotted lines, control staining (primary antibody omitted) of mock infected cells; dashed lines, specific staining of mock infected cells; solid lines, specific staining of retrovirus infected cells. Reprinted with permission from Beerli *et al.* (2000).

away from binding issues and toward application. Of particular interest to gene therapy applications is the ability to regulate activity with a small molecule drug. We have demonstrated one method for achieving this aim, and more work will be directed toward improved methodologies. Previously distant issues such as delivery now need to be addressed; some approaches include viral vectors, facilitated protein uptake, and *ex vivo* gene therapy. Finally, a vast array of site-directed applications are now or will be investigated using fusions to other types of effector domains, such as nucleases (Chandrasegaran and Smith, 1999), integrases (Bushman and Miller, 1997), topoisomerase (Beretta *et al.,* 1999), methylases (Xu and Bestor, 1997), novel enzymes, and protein interaction domains. Thus, polydactyl zinc finger proteins of the future will not only be able to change the flow of trancriptional information in the genome, but they should also be able to permanently modify genomes in very specific ways.

References

Alimandi, M., Romano, A., Curia, M. C. *et al.* (1995). *Oncogene* **10,** 1813–1821.

Al-Lazikani, B., Lesk, A. M., and Chothia, C. (1997). *J. Mol. Biol.* **273,** 927–948.

Altiok, N., Bessereau, J.-L. and Changeux, J.-P. (1995). *EMBO J.* **14,** 4258–4266.

Baca, M., Scanlan, T. S., Stephenson, R. C., and Wells, J. A. (1997). *Proc. Natl. Acad. Sci. USA* **94,** 10063–10068.

Barbas III, C. F., Kang, A. S., Lerner, R. A. and Benkovic, S. J. (1991). *Proc. Natl. Acad. Sci. USA* **88,** 7978–7982.

Barbas III, C. F., Bjorling, E., Chlodi, F. *et al.* (1992). *Proc. Natl. Acad. Sci. USA* **89,** 9339–9343.

Barbas III, C. F., Languino, L. R., and Smith, J. W. (1993a). *Proc. Natl. Acad. Sci. USA* **90,** 10003–10007.

Barbas III, C. F., Collet, T. A., Amberg, W. *et al.* (1993b). *J. Mol. Biol.* **230,** 812–823.

Barbas III, C. F., Hu, D., Dunlop, N. *et al.* (1994). *Proc. Natl. Acad. Sci. USA* **91,** 3809–3813.

Barbas III, C. F. (1995). *Nat. Med.* **1,** 837–839.

Barbas III, C. F., and Burton, D. R. (1996). *Trends in Biotechnology* **14,** 230–234.

Barbas III, C. F., and Wagner, J. (1995). *Methods* **8,** 94–103.

Barbas III, C. F., Heine, A., Zhong, G. *et al.* (1997). *Science* **278,** 2085–2092.

Barbas, C. F., Burton, D. R., Scott, J. K., and Silverman, G. (2000). *Phage Display, A Laboratory Manual,* Cold Spring Harbor Laboratory Press.

Beerli, R. R., Segal, D. J., Dreier, B., and Barbas III, C. F. (1998). *Proc. Natl. Acad. Sci. USA* **95,** 14628–14633.

Beerli, R. R., Dreier, B., Barbas III, C. F. (2000). *Proc. Natl. Acad. Sci. USA* **97,** 1495–1500.

Behrens, A., Schirmer, K., Bols, N. C., and Segner, H. (1998). *Mar. Environ. Res.* **46,** 369–373.

Beretta, G. L., Binaschi, M., Zagni, E., Capuani, L., and Capranico, G. (1999). *Cancer Res* **59,** 3689–3697.

Burton, D. R., Pyati, J., Koduri, R. *et al.* (1994). *Science* **266,** 1024–1027.

Bushman, F. D., and Miller, M. D. (1997). *J. Virol.* **71,** 458–464.

Cao, Y., and Suresh, M. R. (1998). *Bioconjugate Chem.* **9,** 635–644.

Chandrasegaran, S., and Smith, J. (1999). *Biol. Chem.* **380,** 841–848.

Choo, Y., and Klug, A. (1994a). *Proc. Natl. Acad. Sci. USA* **91,** 11163–11167.

Choo, Y., and Klug, A. (1994b). *Proc. Natl. Acad. Sci. USA* **91,** 11168–11172.

Choo, Y., and Klug, A. (1997). *Curr. Opin, Struct, Biol.* **7,** 117–125.

Corbi, N., Libri, V., Fanciulli, M., and Passananti, C. (1998). *Biochem. Biophys. Res. Commun.* **253,** 686–692.

Cowell, I. G. (1994). *Trends Biochem. Sci.* **19,** 38–42.

Desjarlais, J. R., and Berg, J. M. (1992). *Proteins: Struct., Funct., Genet.* **12,** 101–104.

Donato, M. T., Gomez-Lechon, M. J., and Castell, J. V. (1993). *Anal. Biochem.* **213,** 29–33.

Elrod-Erickson, M., Rould, M. A., Nekludova, L., and Pabo, C. O. (1996). *Structure* **4,** 1171–1180.

Elrod-Erickson, M., Benson, T. E., and Pabo, C. O. (1998). *Structure* **6,** 451–464.

Fairall, L., Schwabe, J. W. R., Chapman, L., Finch, J. T., and Rhodes, D. (1993). *Nature* (*London*) **366,** 483–487.

Foote, J., and Winter, G. (1992). *J. Mol. Biol.* **224,** 487–499.

Gottesfeld, J. M., Neely, L., Trauger, J. W., Baird, E. E., and Dervan, P. B. (1997). *Nature* **387,** 202–205.

Greisman, H. A., and Pabo, C. O. (1997). *Science* **275,** 657–661.

Guthrie, J. P., and Wang, J.-P. (1992). *Can. J. Chem.* **70,** 1055–1068.

Hoffmann, T., Zhong, G., List, B. *et al.* (1998). *J. Am. Chem. Soc.* **120,** 2768–2779.

Houbaviy, H. B., Usheva, A., Shenk, T., and Burley, S. K. (1996). *Proc. Natl. Acad. Sci. USA* **93,** 13577–13582.

Hudson, P. J. (1999). *Curr. Opin. Immunol.* **11,** 548–557.

Hynes, N. E., and Stern, D. F. (1994). *Biochim. Biophys. Acta* **1198,** 165–184.

Isalan, M., Choo, Y. and Klug, A. (1997). *Proc. Natl. Acad. Sci. USA* **94,** 5617–5621.

Jamieson, A. C., Kim, S.-H., and Wells, J. A. (1994). *Biochemistry* **33,** 5689–5695.

Jamieson, A. C., Wang, H., and Kim, S.-H. (1996). *Proc. Natl. Acad. Sci. USA* **93,** 12834–12839.

Janda, K. D., Benkovic, S. J., and Lerner, R. A. (1989). *Science* **244,** 437–440.

Jencks, W. P. (1967). In *Catalysis in Chemistry and Enzymology,* McGraw-Hill, New York, p. 288.

Kim, C. A., and Berg, J. M. (1996). *Nat. Struct. Biol.* **3,** 940–945.

Klein, G. and Reymond, J.-L. (1998). *Bioorg. Med. Chem. Lett.* **8,** 1113–1116.

Knappik, A., Ge, L., Honegger, A., Pack, P., Fischer, M., Wellnhofer, G., Hoess, A., Wolle, J. Pluckthun A., and Virnekas, B. (2000). *J. Mol. Biol.* **296,** 57–86.

Kohler, G., and Milstein, C. (1975). *Nature* **256,** 495–497.

Kraus, M. H., Fedi, P., Starks, V., Muraro, R., and Aaronson, S. A. (1993). *Proc. Natl. Acad. Sci. USA* **90,** 2900–2904.

Krizek, B. A., Amann, B. T., Kilfoil, V. J., Merkle, D. L., and Berg, J. M. (1991). *J. Am. Chem. Soc.* **113,** 4518–4523.

Lai, C. Y., Nakai, N., and Chang, D. (1974). *Science* **183,** 1204–1206.

Lee, M. S., Gippert, G. P., Soman, K. V., Case, D. A., and Wright, P. E. (1989). *Science* **245,** 635–637.

Lee, K.-F., Simon, H., Chen, H. *et al.* (1995). *Nature* **378,** 394–398.

List, B., Shabat, D., Barbas III, C. F., and Lerner, R. A. (1998a). *Chem. Eur. J.* **4,** 881–885.

List, B., Barbas III, C. F. and Lerner, R. A. (1998b). *Proc. Natl. Acad. Sci. USA* **95,** 15351–15355.

List, B., Shabat, D., Zhong, G. *et al.* (1999). *J. Am. Chem. Soc.* **121,** 7283–7291.

Liu, Q., Segal, D. J., Ghiara, J. B., and Barbas III, C. F. (1997). *Proc. Natl. Acad. Sci. USA* **94,** 5525–5530.

Lo, C.-H. L., Wentworth, J., P., Jung, K. W. *et al.* (1997). *J. Am. Chem. Soc.* **119,** 10251–10252.

Margolin, J. F., Friedman, J. R., Meyer, W. K.-H. *et al.* (1994). *Proc. Natl. Acad. Sci. USA* **91,** 4509–4513.

Matteucci, M. D., and Wagner, R. W. (1996). *Nature* **384,** 20–22.

Miller, J., McLachlan, A. D., and Klug, A. (1985). *Embo. J.* **4,** 1609–1614.

Morea, V., Tramontano, A., Rustici, M., Chothia, C., and Lesk, A. M. (1998). *J. Mol. Biol.* **275,** 269–294.

Morris, A. J., and Tolan, D. R. (1994). *Biochemistry* **33,** 12291–12297.

Narayan, V. A., Kriwacki, R. W., and Caradonna, J. P. (1997). *J. Biol. Chem.* **272,** 7801–7809.

Niculesco-Duvaz, I., and Springer, C. J. (1997). *Adv. Drug Delivery Rev.* **26,** 151–172.

Nolte, R. T., Conlin, R. M., Harrison, S. C., and Brown, R. S. (1998). *Proc. Natl. Acad. Sci. USA* **95,** 2938–2943.

Padlan, E. A. (1994). *Mol. Immunol.* **31,** 169–217.

Pauling, L. (1946). *Chem. Eng. News* **24,** 1375.

Pauling, L. (1947). In *Silliman Lecture,* Yale University Press, New Haven.

Pavletich, N. P., and Pabo, C. O. (1991). *Science* **252,** 809–817.

Pavletich, N. P., and Pabo, C. O. (1993). *Science* (Washington. D. C., 1883–) **261,** 1701–1707.

Pini, A., Viti, F., Santucci, A., Carnemolla, B., Zardi, L., Neri, P., Neri, D. (1998). *J. Biol. Chem.* **273,** 21769–21776.

Pollack, S. J., Hsiun, P., and Schultz, P. G. (1989). *J. Am. Chem. Soc.* **111,** 5961–5962.

Ptashne, M. (1997). *Nature Medicine* **3,** 1069–1072.

Rader, C., and Barbas III, C. F. (1997). *Curr. Opin. Biotechnol.* **8,** 503–508.

Rader, C., and List, B. (2000). *Chem. Eur. J.* **6,** 2091–2095.

Rader, C., Cheresh, D. A., and Barbas III, C. F. (1998). *Proc. Natl. Acad. Sci. USA* **95,** 8910–8915.

Rader, C., Ritter, G., Nathan, S. *et al.* (2000). *J. Biol. Chem.* **275,** 13668–13676.

Rader, C., and Barbas III, C. F. (2000). In *Phage Display, A Laboratory Manual* (ed. Barbas, C. F. *et al.*), Cold Spring Harbor Laboratory Press.

Rebar, E. J., and Pabo, C. O. (1994). *Science* (Washington, D.C., 1883–) **263,** 671–673.

Rhodes, D., and Klug, A. (1993). *Scientific American* **268,** 56–59.

Riechmann, L., Clark, M., Waldmann, H., and Winter, G. (1988). *Nature* **332** (6162), 323–327.

Rosok, M. J., Yelton, D. E., Harris, L. J. *et al.* (1996). *J. Biol. Chem.* **271**(37), 22611–22618.

Sadowski, I., Ma, J., Triezenberg, S., and Ptashne, M. (1988). *Nature* **335,** 563–564.

Saito, S., Hatanaka, K., Kano, T. and Yamamoto, H. (1999). *Angew. Chem. Int. Ed.* **37,** 3378–3381.

Sanz, I. (1991). *J. Immunol.* **147,** 1720–1729.

Segal, D. J., Dreier, B., Beerli, R. R., and Barbas III, C. F. (1999). *Proc. Natl. Acad. Sci. USA* **96**(6), 2758–2763.

Segal, D. J., and Barbas III, C. F. (2000). *Curr. Opin. Chem. Biol.* **4**(1), 34–39.

Schier, R., McCall, A., Adams, G. P. *et al.* (1996). *J. Mol. Biol.* **263**(4), 551–567.

Shabat, D., Rader, C., List, B., Lerner, R. A., and Barbas III, C. F. (1999). *Proc. Natl. Acad. Sci. USA* **96,** 6925–6930.

Siegel, P. M., Ryan, E. D., Cardiff, R. D., and Muller, W. J. (1999). *Embo J.* **18,** 2149–2164.

Sinha, S. C., Barbas III, C. F., and Lerner, R. A. (1998). *Proc. Natl. Acad. Sci. USA* **95,** 14603–14608.

Smith, J. W., Hu, D., Satterthwait, A. C., Pinz-Sweeney, S. and Barbas III, C. F. (1994). *J. Biol. Chem.* **269,** 32788–32795.

Steinberger, P., Andris-Widhopf, J., Buhler, B., Torbett, B. E., and Barbas III, C. F. (2000). *Proc. Natl. Acad. Sci. USA* **97**(2), 805–810.

Suzuki, M., and Yagi, N. (1994). *Proc. Natl. Acad. Sci. USA* **91,** 12357–12361.

Tanaka, F., Kinoshita, K., Tanimura, R., and Fujii, I. (1996). *J. Am. Chem. Soc.* **118,** 2332–2339.

Tramontano, A., Janda, K. D., and Lerner, R. A. (1988). *J. Am. Chem. Soc.* **110,** 2282–2286.

Turner, J. M., Bui, T., Lerner, R. A., Barbas III, C. F., and List B. (2000). *Chem. Eur. J.* in press.

Vaughan, T. J., Osbourn, J. K., and Tempest, P. R. (1998). *Nat. Biotechnol.* **16**(6), 535–539.

Wade, H., and Scanlan, T. S. (1996). *J. Am. Chem. Soc.* **118,** 6510–6511.

Wagner, J., Lerner, R. A., and Barbas III, C. F. (1995). *Science* **270,** 1797–1880.

Wolfe, S. A., Greisman, H. A., Ramm, E. I., and Pabo, C. O. (1999). *J. Mol. Biol.* **285**(5), 1917–1934.

Wu, H., Yang, W.-P., and Barbas III, C. F. (1995). *PNAS* **92,** 344–348.

Wu, T. T., Johnson, G., and Kabat, E. A. (1993). *Proteins* **16,** 1–7.

Wuttke, D. S., Foster, M. P., Case, D. A., Gottesfeld, J. M., and Wright, P. E. (1997). *J. Mol. Biol.* **273**(1), 183–206.

Xu, G.-L., and Bestor, T. H. (1997). *Nature Gen.* **17,** 376–378.

Yang, W.-P., Green, K., Pinz-Sweeney, S. *et al.* (1995). *J. Mol. Biol.* **254,** 392–403.

Zaug, A. J., Been, M. D., and Cech, T. R. (1986). *Nature* **324,** 429–433.

Zhong, G., Hoffmann, T., Lerner, R. A., Danishefsky, S., and Barbas III, C. F. (1997). *J. Am. Chem. Soc.* **119,** 8131–8132.

Zhong, G., Shabat, D., List, B. *et al.* (1998). *Angew. Chem. Int. Ed.* **37,** 2481–2484.

Zhong, G., Lerner, R. A., and Barbas III, C. F. (1999). *Angew. Chem.* **111,** 3957–3960.

IN VITRO SELECTION AND EVOLUTION OF PROTEINS

By ANDREAS PLÜCKTHUN, CHRISTIANE SCHAFFITZEL, JOZEF HANES, and LUTZ JERMUTUS

Biochemisches Institut, Universität Zürich, Switzerland

I. Introduction

Since the advent of recombinant DNA technology, the engineering of proteins for improved binding specificities, ligand affinities, and stability has become commonplace. Both hypothesis-based ''rational'' and combinatorial approaches exist to address these tasks. It is the technical advances in the latter and the apparent current limitations in the former which gave rise to this volume.

To rationally (re)design a protein requires detailed structural and, in the case of enzymes, mechanistic information. The technical problem of producing the gene for virtually any protein with any sequence has been solved by gene synthesis and site-directed mutagenesis, but the knowledge of how sequence changes affect protein expression, function and biophysical properties lags far behind. Currently, the predictive accuracy of even the most sophisticated structure-based engineering approaches is often still insufficient to produce the desired design effects without additional experimentation (Dougan *et al.,* 1998; Yelton *et al.,* 1995). Nevertheless, progress has been achieved and is certain to continue (Rubingh, 1997; Hellinga, 1997). The fundamental problem of

0065-3233/00 $35.00

predictions is the multitude of configurations of very similar energy. It is very difficult to decide which way the balance will tip and to predict whether a large number of possible small movements will or will not result in a larger overall conformational change. If no structural and mechanistic information is available, the effects of mutational changes are almost fully unpredictable.

In contrast, for combinatorial approaches the challenge lies in the technology—to actually achieve a Darwinian evolution in a reasonable time. Although the global concepts have been clear for a long time, only recently have the tools become available to exploit this approach in practice. Currently, it appears that combinatorial and evolutionary methods are in the lead for actually improving a particular molecule in practice.

It is useful at this point to define the differences between combinatorial and evolutionary strategies. The underlying principle of the *combinatorial* approach is selection for the desired property from a pool of diverse molecules (a single-pot or ''constant'' library, which does not change later). In this case, the accessed sequence space equals the functional library size. In the *evolutionary* approach, on the other hand, a given molecule (one starting sequence) or a library (many starting sequences) is continuously diversified, to elicit improved or even entirely novel functions by an iterative process of diversification and selection. Therefore, the accessed sequence space in an evolution experiment is far greater than the initial library size.

Directed evolution mimics the natural process by which protein variants arise and are tested for their fitness in living systems in a series of ''generations''—cycles of diversification and selection. A particularly instructive example (which occurs much more rapidly than the phylogenetic evolution of proteins) is the somatic hypermutation of antibodies. During the secondary immune response, antibody V genes undergo point mutations at a frequency of about 10^{-3} per base pair per generation (Berek and Milstein, 1987; Allen *et al.,* 1988). Concomitant with this rapid mutational process, selection of B cells with high-affinity receptors for the immunizing antigen leads to a 10- to 100-fold increase in the average antibody affinity (Berek and Milstein, 1987). This ''affinity maturation'' can be modeled as an adaptive walk on a rugged sequence landscape (Macken and Perelson, 1989), and it was found that a small number of single mutations is necessary to reach a local optimum in the fitness landscape.

To date, a number of well-established strategies exist to select protein-ligand interactions. These can be exploited to *identify* binding molecules (combinatorial approach) or iteratively *improve* them (evolutionary approach). Additionally, if the binding interaction is restricted to the native state, the selection can be used to select for the quality of the protein, including expression yield (Jung *et al.,* 1999), folding kinetics and thermodynamic stability (Ruan *et al.,* 1998; Spada *et al.,* 1998), resistance to proteolysis (Sieber *et al.,* 1998; Kristensen and Winter, 1998) or stability in an extreme nonphysiological environment (Kuchner and Arnold, 1997; Jung *et al.,* 1999; Schmidt-Dannert and Arnold, 1999).

Examples of such protein-protein interaction selection systems are phage display (Smith, 1985; Winter *et al.,* 1994), display on other viruses (Kasahara *et al.,* 1994), bacterial surface display (Georgiou *et al.,* 1993; Daugherty *et al.,* 1999), yeast display (Kieke *et al.,* 1997; Boder and Wittrup, 1997), the yeast two hybrid system (Fields and Song, 1989; Chein *et al.,* 1991), and protein-fragment complementation assays (Pelletier *et al.,* 1998). These methods all contain a necessary *in vivo* step, which has a number of disadvantages that will be discussed in the following sections.

Ribosome display (Hanes and Plückthun, 1997) is the first method for screening and selecting functional proteins that is performed entirely *in vitro,* thus circumventing many of the drawbacks of *in vivo* systems. Here, we present the principles underlying ribosome display and some of its applications for generating high affinity and high stability antibodies from given starting molecules of complex libraries and summarize related *in vitro* selection technologies. We also compare *in vitro* selection to *in vivo* methods.

In ribosome display, the physical link between genotype and phenotype is accomplished by mRNA–ribosome–protein complexes, which are directly used for selection. If a library of different mRNA molecules is translated, a protein library results in which each protein is produced from its ''own'' mRNA and remains connected to it. Since these complexes of the proteins and their encoding mRNAs are stable for several days under the appropriate conditions, very stringent selections can be performed. As all steps of ribosome display are carried out *in vitro,* reaction conditions of the individual steps can be tailored to the requirements of the protein species investigated, as well as the objectives of the selection or evolution experiment. Application of ribosome display has produced scFv fragments of antibodies with affinities in the picomolar range from libraries prepared from immunized mice (Hanes *et al.,* 1998) and more recently from a naive, completely synthetic library (Hanes *et al.,* 2000), and has been used to evolve improved off-rates and stability (Jermutus *et al.,* 2000).

A. In Vivo *versus* In Vitro *Selections*

All *in vivo* methods have in common that the library usually encoded on a plasmid or phage replicon must be transformed into cells, either bacteria or yeast. These microorganisms then express the protein for an intracellular interaction screen, such as the yeast two hybrid system (Fields and Song, 1989; Chien *et al.,* 1991) or the protein-fragment complementation assay (Pelletier *et al.,* 1998; 1999). Alternatively, bacteria or yeast cells display the protein on their surface (Georgiou *et al.,* 1993; Daugherty *et al.,* 1999; Kieke *et al.,* 1997; Boder and Wittrup, 1997). Finally, the bacteria may be transformed with the library in order to produce phages (e.g., phage display with filamentous or λ phages) that then carry the protein on their surface. Obviously, the library size is determined by the transformation frequency, and typically *Escherichia coli* libraries of 10^{10} to 10^{11} present an upper limit (Dower and Cwirla, 1992). To create libraries even with this size involves significant labor. Importantly, after each *in vitro* randomization step, a new library has to be created and transformed. Libraries screened by yeast display (Boder and Wittrup, 1997) and the yeast two hybrid system (Fields and Song, 1989) are even smaller, due to the generally lower transformation efficiency in yeast.

With ribosome display and other *in vitro* selection systems (see below) no transformation is necessary. Therefore, it is possible to assemble libraries *in vitro* and retain their very large size. Furthermore, it is possible to screen these protein and peptide libraries with 10^{11} or more members, with a new library of point mutants at every generation in an evolution experiment. An increase in library size improves the chance to select for the desired function and in addition increases the diversity of molecules selected. Lancet *et al.* (1993) estimated the relationship between the library size and the best affinity of a member in the library that could be selected. The prediction was that increasing a library from 10^8 to 10^{12} sequences will increase the affinity of the best selected binder up to 300-fold. Therefore, with the possibility of screening very large libraries by *in vitro* selection technology, it becomes more likely that a larger variety of high affinity binders with the desired function are selected. Using ribosome display, it has indeed been possible to select and evolve high affinity antibodies with dissociation constants as low as 80 pM from protein libraries (Hanes *et al.,* 1998; Hanes *et al.,* 2000) in a short time. In this approach, very large libraries are easily accessible because they do not have to be cloned and transformed into cells but can be rapidly assembled *in vitro.*

In vivo, a pre-selection due to the host environment cannot be avoided. Growth disadvantage or even toxicity can lead to a loss of potential candidates. More rapidly growing library members can become overrepresented in the culture despite the fact that they are not specifically enriched by the actual selection. Furthermore, folding, transport, aggregation, and proteolytic degradation can often not be controlled effectively by the investigator in an *in vivo* environment, and the final application of the selected molecule may be envisioned for an environment quite different from that of *in vivo* selection. Another important point is that cells are complex genetic entities, and they often find many ways to survive or evade the selection pressure that differ from the ones desired by the investigator. Mutations or recombinations within the plasmid or within the host genome can provide an easy solution for the cell to circumvent the selection pressure.

These undesired selection pressures are substantially reduced *in vitro,* and the translation conditions can be optimized for the protein to be displayed on a case-by-case basis. Ribosome display can also be easily combined with *in vitro* mutagenesis techniques such as mutagenic PCR (Cadwell and Joyce, 1992), DNA shuffling (Stemmer, 1994), the staggered extension process (Zhao *et al.,* 1998) or other recombination-based methods in an evolution experiment (see Section IV, B). Also, if nonproofreading DNA polymerases are used for ribosome display, a diversification of the initial library during the selection cycles will be observed automatically due to mutations introduced during the many PCR steps at the end of each selection cycle. Thereby, the sequence space sampled is much larger than the initial size of the library. In principle, the quality of the pool is iteratively improved, since only proteins that survived the first selection will be used for further diversification. During all subsequent selections the mutated proteins have to compete with their progenitors.

In contrast, if a diversification step needs to be included in an *in vivo* selection strategy in order to evolve the protein under investigation, either a mutator strain needs to be used (Low *et al.,* 1996) or it is necessary to repeatedly switch between the selection procedure *in vivo* (phage, bacteria, yeast) and the mutagenesis step for diversification carried out *in vitro.* The disadvantage of the former case is that mutator strains can also create unwanted mutations in the plasmid and in the host genome, while the latter is a rather laborious procedure, as after each diversification step, the newly created library has to be religated and retransformed. Consequently, only relatively few examples of protein evolution over several cycles of diversification and selection are found in the literature (e.g., Yang *et al.,* 1995; Schier and Marks, 1996; Moore *et al.,* 1997).

Ribosome display has thus two main advantages compared to *in vivo* selection systems: on the one hand *in vitro* technologies allow one to screen very large libraries, since no transformation steps are necessary. On the other hand, subsequent diversification of the library is easy and convenient and every single clone present in the library can conceivably be evolved. In addition, working *in vitro* allows for tight control of the selection experiment at each step.

II. The Key to *In Vitro* Protein Evolution: Cell-free Translation

A basic understanding of *in vitro* translation is a prerequisite for devising and optimizing cell-free protein selection systems. *In vitro* protein synthesis, independent of its use in selection technology, has received increasing interest in recent years (reviewed by Jermutus *et al.,* 1998). This is due to improvement of protein yields, the increase in detection sensitivity of many analytical methods, and the advent of new technologies such as atomic force microscopy (AFM) (Engel *et al.,* 1999) and fluorescence correlation spectroscopy (FCS) (Eigen and Rigler, 1994; Rigler, 1995) that allow analytical work at low protein concentrations, even down to the single molecule level.

Many applications of cell-free translation rely on the correct folding of the *in vitro* expressed polypeptide into its three-dimensional structure, and this is a prerequisite for all protein selection systems that are based on *in vitro* translation. Because proteins are selected and evolved for functionality, sufficient expression and correct folding in the respective cell-free translation system are a necessity for efficient selection. An attractive advantage of using *in vitro* translations is that, at least in principle, any component of the reaction can be deliberately added or removed. To achieve any improvement in yield, however, separate consideration of both the actual translation and the folding is necessary. Even in optimized systems, however, translation yields are not similar for all globular proteins.

A. *Increasing* In Vitro *Translation Yields*

There are many hypotheses about the underlying mechanisms of differences in translation yields, and to date no cell-free translation system has been engineered that allows a high expression and quantitative folding of any given protein sequence. At least three problems need to be solved regarding total protein production. First, mRNA secondary structures can inhibit translation initiation or stall elongating ribosomes (Kozak, 1989; Yu *et al.,* 1994). This becomes especially important if RNA hairpin loops are further stabilized in nonphysiological conditions such as the relatively high Mg^{2+} concentrations typically used in a standard S30 *E. coli* translation system. However, using the degeneracy of the genetic code, this limitation can be removed by silent mutagenesis of the primary sequence. Second, RNase and protease sensitivity can reduce expression yield by removing the template RNA or the synthesized protein (see

Section III, B, 2 for discussion of *E. coli* RNases). Because the recognition sequences of these enzymes are different in each organism and are in many cases unknown, this problem might be solved by removing these enzymes from the translation system with immunoprecipitation.

A third and more general bottleneck is tRNA availability. Any cell-free expression system contains endogenous aminoacyl-tRNA-synthetases, and usually a tRNA pool from the same organism is added for the translation reaction. Because the genetic code is degenerate, the pool contains tRNAs with different anti-codons for the same amino acid. The concentrations for these different tRNAs vary, resulting in rare codons on the mRNA level. Although some reports have suggested that these codons might be important for cotranslational folding (Komer *et al.,* 1999; Thanaraj and Argos, 1996), this point remains controversial and inclusion of rare codons generally decreases protein yield. Ribosome stalling at rare codons can either trigger 10Sa-RNA-mediated proteolytic degradation (Roche and Sauer, 1999) or premature translation termination (Komar *et al.,* 1999). Simply increasing the total concentration of the whole tRNA pool would not change the molar ratios of the different tRNAs that are competing for aminoacylation at the synthetase. As a consequence, codon rarity will persist. The only way to resolve the problem lies in adding a tRNA pool with different molar rations of the tRNAs accepting the same amino acid (De Pasquale and Kanduc, 1998). *In vivo,* protein translation was found to be mediated by changes of the tRNA pool composition (Kanduc, 1997; Hentze, 1995). *In vitro* transcribed tRNAs, which can be obtained with T7 RNA polymerase, are substrates for aminoacyl-tRNA-synthetases *in vitro.* This opens the door for the rational design of the tRNA pool to be added in cell-free translation.

Moreover, it was recently shown that the deliberate removal of phosphatases from the *in vitro* translation mixture by immunoprecipitation results in increased protein yields (Shen *et al.,* 1998). As proposed earlier (Jermutus *et al.,* 1998), the elimination of one of the many causes of fast ATP and GTP depletion extends the time of synthesis, and, as a consequence, the total amount of produced protein. It should be possible to similarly remove proteases and nucleases (see above). Together with new ATP regeneration systems to keep the biochemical energy level at a high steady state (Kim and Swartz, 1999), the optimization measures mentioned should increase the level of synthesis of most proteins, and this should directly improve *in vitro* selection.

B. Increasing the Fraction of Functional Molecules

Cotranslational misfolding or aggregation reduces the active fraction of synthesized proteins and might also decrease the expression yield by inhibiting the translating ribosome. However, correct folding can at present only be improved on a case-to-case basis, even though the use of standardized cocktails of beneficial factors is conceivable. The optimization of cell-free production of single-chain Fv antibody fragments (scFv) has been investigated in detail in an *E. coli* translation system (Ryabova *et al.,* 1997). The factors contributing to efficient folding and proper disulfide bond formation were identified. It was suggested that chaperones, mostly DnaK and DnaJ, increase the amount of soluble scFv in an *E. coli* S30 cell-free system, but do not affect the amount of functional proteins, indicating that there are soluble, misfolded species. Eukaryotic PDI (protein disulfide isomerase), a eukaryotic protein catalyzing disulfide bond formation and rearrangement, added cotranslationally in the cell-free system, increased the amount of functional antibodies. Evidence was provided that the isomerization reaction, and not the net disulfide bond formation, is the rate-limiting step for the *in vitro* folding of scFv fragments. The *in vitro* production of disulfide-containing, native molecules is thus possible, and the use of appropriate redox conditions and addition of folding catalysts can have a significant effect on the preparative production of biologically active proteins *in vitro.* However, different proteins may need other chaperone cocktails to maximize functionality.

III. *In Vitro* Selection Strategies

In the following sections, an overview of the different approaches for *in vitro* selection is provided. Although this chapter focuses on proteins, we want to briefly explain a nucleic acid selection method that formed the basis of the *in vitro* evolution of functional proteins.

A. SELEX

In 1990, Tuerk and Gold introduced a technology termed *S*ystematic *E*volution of *L*igands by *EX*ponential Enrichment (SELEX). By using SELEX, it is possible to exponentially enrich and evolve RNA ligands in multiple rounds *in vitro* from a random oligonucleotide pool. Many protein motifs and functions can be mimicked by folded RNA structures (Roberts and Ja, 1999). Currently, this method is widely employed to screen for nucleic acid ligands (aptamers) binding to numerous targets with potential applications in diagnostics and biotechnology (Osborne *et al.,* 1997). Furthermore, SELEX has been used to isolate novel nucleic acid based catalysts for a variety of reactions (Gold *et al.,* 1995).

In SELEX, multiple rounds of *in vitro* transcription of random nucleic acid pools, affinity selection, and RT-PCR are performed, thus giving rise to exponential amplification of the selected molecules. The principle underlying SELEX is schematically depicted in Figure 1. After several selection cycles, the binders can subsequently be cloned and sequenced and then characterized. In SELEX, genotype and phenotype are simultaneously represented by the same RNA molecule, since it exerts its function through its three-dimensional structure, which is in turn determined by its nucleotide sequence. The chemical and functional diversity of RNA can be further increased by addition of cofactors such as histidine (Roth and Breaker, 1998) and divalent cations (Tarasow *et al.,* 1997) to the selection.

Nevertheless, RNA molecules have some severe disadvantages as ligands, and they have been almost completely replaced by peptides and proteins in evolution. RNA is a polyanion and thus frequently selects positive charges as the target (Hermann and Patel, 2000), thereby restricting the epitopes on a protein that can be blocked. Furthermore, RNA is extremely prone to degradation by ubiquitous RNases. To actually use an aptamer in any application, the RNA has first to be stabilized by introducing stable nucleotide analogs that are obtained either by the

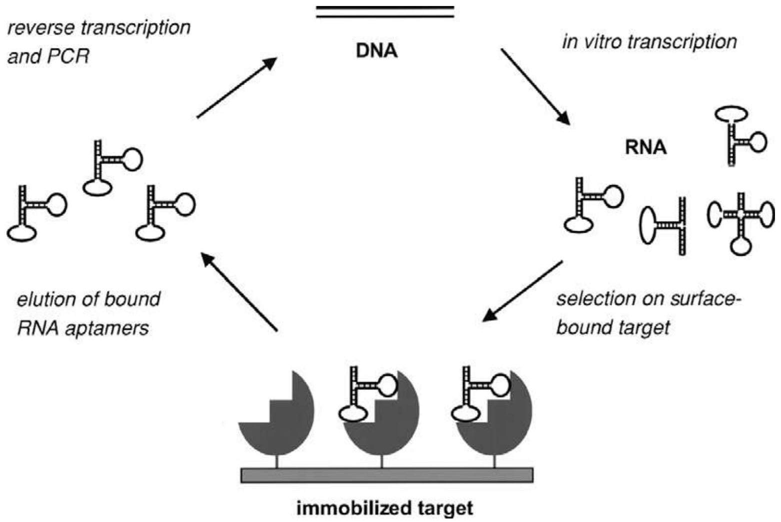

reverse transcription and PCR
DNA
in vitro transcription
RNA
elution of bound RNA aptamers
selection on surface-bound target
immobilized target

FIG. 1. SELEX. A DNA oligonucleotide pool is transcribed *in vitro.* The resulting RNA is directly used in affinity selection against an immobilized target. RNA molecules that bind (termed ''aptamers'') are subsequently eluted. By RT-PCR, an oligonucleotide pool enriched for binders can be regenerated and used for a new round of SELEX.

addition of phosphorothioates or the substitution of 2′-OH groups by 2′-NH_2 and 2′-F (Eaton *et al.,* 1997; Ruckman *et al.,* 1998). Such molecules cannot be synthesized by enzymes in preparative amounts, but must be prepared by large-scale synthesis. The initial (and remaining) appeal of SELEX was the very rapid generation of high affinity binders from very large initial libraries (10^{15} to 10^{16} sequences), as the RNA transcripts directly constitute the ligands.

B. Ribosome Display

In their original publication about SELEX, Tuerk and Gold (1990) already speculated that a similar approach could be adapted to protein selection. They referred to experiments describing the isolation of particular mRNAs from a pool of variants by immunoprecipitation of the nascent polypeptides present in the mRNA-ribosome-polypeptide complexes (Korman *et al.,* 1982; Kraus and Rosenberg, 1982). In fact, soon after the publication of SELEX (Tuerk and Gold, 1990) a patent application was filed (Kawasaki, 1991), proposing a similar approach to enrich peptides from libraries.

The first experimental demonstration of the ribosome display technology was the selection of short peptides from a library using an *E. coli* S30 *in vitro* translation system (Mattheakis *et al.,* 1994; Mattheakis *et al.,* 1996). The concept pursued by Mattheakis and coworkers (1994) for peptides was then used for the development of ribosome display of functional proteins by use of the *E. coli* S30 *in vitro* translation system (Hanes and Plückthun, 1997). However, for this purpose it was necessary to significantly modify and optimize the experimental conditions of ribosome display to make this technology efficient enough for the display of correctly folded, functional proteins.

1. Principle of Ribosome Display

The principle of ribosome display is depicted in Figure 2. A DNA library, encoding a polypeptide in a special ribosome display cassette (discussed in Section III, B, 2) is either directly used for coupled *in vitro* transcription-translation, or first transcribed *in vitro* to mRNA, purified, and subsequently used for *in vitro* translation (discussed in Section III, B, 3). During *in vitro* translation ribosomal complexes (mRNA-ribosome-polypeptide) form that contain a functionally folded protein emerging from the ribosomal tunnel and most probably still connected to the tRNA at its C-terminal end. For the protein to fold, it must have an unstructured region occupying the ribosomal tunnel, which must be encoded downstream of the gene region that encodes the folded protein (see Section III, B, 2). The composition and reaction conditions of the

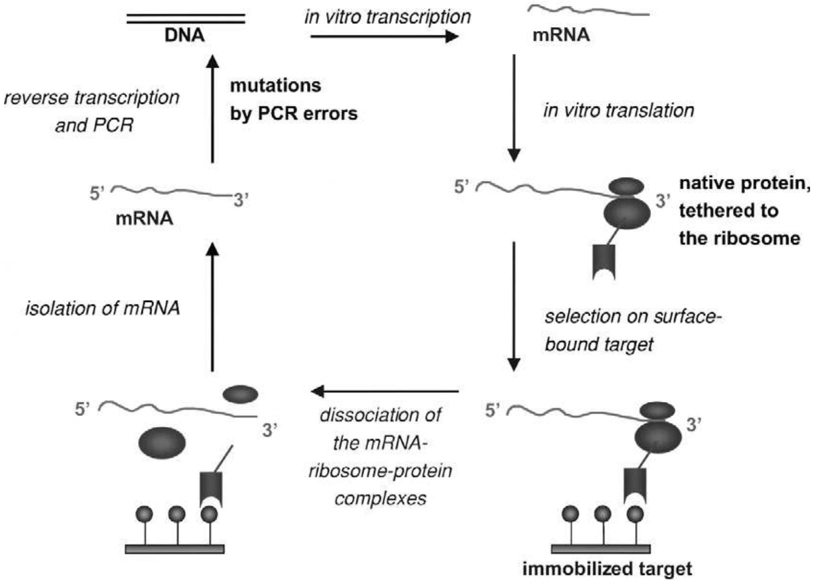


FIG. 2. Ribosome display. A library of proteins (e.g., scFv fragments of antibodies) is transcribed and translated *in vitro*. The resulting mRNA lacks a stop codon, giving rise to linked mRNA-ribosome-protein complexes. These are directly used for selection on the immobilized target. The mRNA incorporated in bound complexes is eluted and purified. RT-PCR can introduce mutations and yields a DNA pool enriched for binders that can be used for the next iteration.

in vitro translation must be commensurate with folding of the particular protein in question (see Section III, B, 3).

In the *E. coli* system, it is important to stop the *in vitro* translation reaction by rapid cooling on ice. The reaction is usually diluted severalfold in prechilled buffer containing the components for stabilization of the ribosomal complexes. In the *E. coli* system, the ribosomal complexes can be very efficiently stabilized by low temperature and by high Mg^{2+} concentrations (50 mM), and then used for affinity selection. It is believed that high Mg^{2+} ''condenses'' the ribosome by binding to the rRNA, making it difficult for the peptidyl-tRNA to dissociate or be hydrolyzed. The low temperature probably slows down the hydrolysis of the peptidyl-tRNA ester bond, and perhaps also the thermal motions, which would facilitate dissociation of the peptidyl-tRNA. Such complexes are stable for up to several days.

a. Binding Selection. In principle, the affinity selection of the ribosomal complexes can be performed in two different ways. Either ligands can be immobilized on a surface (panning tubes or microtiter wells, for instance), or biotinylated ligands can be used, which bind to the proteins displayed on the ribosomal complexes and are subsequently captured by streptavidin-coated magnetic beads. Usually, the panning is performed for one hour at temperatures equal to or below 4°C for routine enrichments. However, shorter as well as longer incubation times (as long as twenty days) are possible and can be advantageous. The former can be useful if all binders are to be captured, regardless of their affinity, while the latter is appropriate if very high affinity binders are to be evolved (see Section IV, B, 2). The advantage of performing the panning in solution is that the ligand concentration is well defined and the ligand is mostly in its native conformation. Therefore, the number of unspecifically bound ribosomal complexes is usually lower. Furthermore, nonspecific binding during affinity selection of ribosomal complexes can be decreased using diluted, autoclaved milk and heparin (Hanes *et al.*, 1998). Heparin may additionally act also as an RNase inhibitor.

b. Elution. After affinity selection, nonspecifically bound ribosomal complexes are removed by intensive washing with magnesium-containing buffer. Removal of the stabilizing Mg^{2+} ions with an excess of EDTA causes dissociation of all bound complexes, allowing the mRNA of bound ribosomal complexes to be directly isolated and obviating the need to elute the binder from the target. Alternatively, by competitive elution of bound ribosomal complexes with free ligand, followed by mRNA isolation of eluted complexes, only the mRNA present in complexes containing a functional and specific binding protein is isolated, possibly leading to higher enrichment factors. However, this approach might be difficult to apply for binders with a very high affinity for antigen.

c. Amplification. The isolated mRNA is then used for RT-PCR, and the amplified DNA can be used for the next cycle of ribosome display. When magnetic beads are used for selection, RT-PCR may also be performed directly with the washed beads (He and Taussig, 1997). After each round of ribosome display, a portion of the DNA can be analyzed by cloning and sequencing and by ELISA or RIA.

2. *The Ribosome Display Construct*

The features of the ribosome display construct are summarized in Figure 3. On the DNA level, the construct requires a T7 promoter for efficient *in vitro* transcription to mRNA. On the mRNA level, the construct contains, as a regulatory sequence for translation, either a prokaryotic ribosome binding site (Shine and Dalgarno, 1975) if the *E. coli*

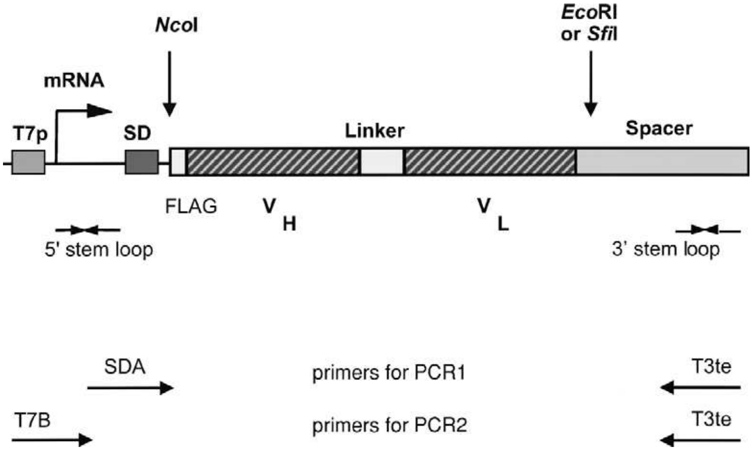


FIG. 3. DNA construct used for *E. coli* ribosome display, as illustrated for a single-chain antibody. The promoter (T7) is followed by a Shine-Dalgarno sequence (SD) and the protein of interest (here a scFv construct) containing an N-terminal FLAG tag for detection. The variable domains V_H and V_L are joined by a glycine/serine-rich linker. A spacer (tether) is cloned in frame behind the sequence of the antibody scFv fragment without a stop codon. Sequences encoding RNA stem-loop structures are present both at the 5′ and 3′ ends. In ribosome display, after the reverse transcription step (with primer T3te), two subsequent PCR steps are used to reintroduce the Shine-Dalgarno sequence (PCR1; primers SDA, T3te) and the T7 promoter (PCR2; primers T7B, T3te) to regenerate the complete scFv construct.

system is used, or a Kozak consensus and enhancer sequence (Kozak, 1984) if the eukaryotic ribosome display system is used. This sequence is followed by the open reading frame encoding the protein to be displayed, followed by a spacer sequence fused in frame to the protein.

The coding region ends with the protein sequence—that is, there is no stop codon present. In the prokaryotic system the presence of a stop codon would result in the binding of the release factors (Grentzmann *et al.,* 1995; Tuite and Stansfield, 1994) and the ribosome recycling factor (Janosi *et al.,* 1994) to the mRNA-ribosome-protein complexes. This would then lead to the release of the protein by hydrolysis of the peptidyl-tRNA (Tate and Brown, 1992), thereby dissociating the ribosomal complexes (Fig. 4A). A similar mechanism exists in eukaryotic systems (Frolova *et al.,* 1994; Zhouravleva *et al.,* 1995).

Another important prerequisite for efficient ribosome display in the *E. coli* system is the elimination of the 10Sa-RNA (Ray and Apirion, 1979). 10Sa-RNA is a stable bacterial RNA with a tRNA-like structure, but having an extended Ω-loop (Komine *et al.,* 1994). If a truncated

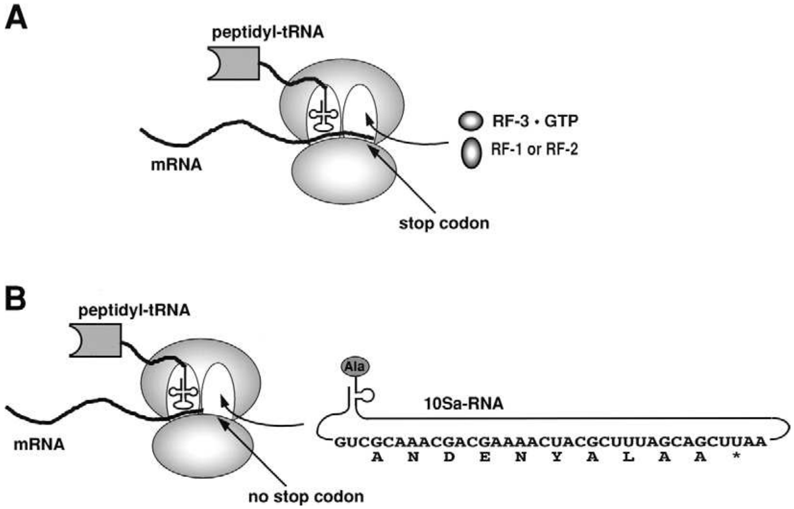


FIG. 4. Role of the stop codon and 10Sa-RNA in *E. coli* translation. (A) When a stop codon is encountered, a complex of two release factors, RF-1 and RF-3 or RF-2 and RF-3, binds instead of the tRNA. The release factor RF-1 recognizes the stop codons UAA and UAG, while RF-2 recognizes UAA and UGA. The binding of the release factor complex results in hydrolysis of the peptidyl-tRNA and release of the peptide. (B) The role of 10Sa-RNA. If truncated mRNA without a stop codon is translated in *E. coli*, the ribosome stops at the end of the mRNA. 10Sa-RNA can then bind to the ribosomal A site and 10Sa-RNA can act as tRNA by transferring an alanine to the truncated protein. Subsequently, 10Sa-RNA acts as mRNA and a peptide tag with the indicated sequence is added to the truncated protein. 10Sa-RNA encodes a stop codon and therefore the protein is released and then degraded by proteases specifically recognizing this C-terminal tag.

mRNA, lacking a stop codon, is translated *in vivo* in *E. coli,* the 10Sa-RNA binds to the ribosomal tRNA acceptor site. This results in a carboxy-terminal modification of the truncated polypeptide by addition of a peptide tag encoded by the 10Sa-RNA and subsequent release from the ribosome (Fig. 4B). The released protein, tagged with this sequence, is finally degraded by a tail-specific protease (Keiler *et al.,* 1996; Roche and Sauer, 1999).

At both ends of the mRNA, the ribosome display construct should include stemloops, 5′- and 3′-stemloops are known to stabilize mRNA against RNases *in vivo* as well as *in vitro.* The presence of stemloops is important, especially in the *E. coli* ribosome display system, because the extract used for *in vitro* translation contains high RNase activities. To date, five of twenty *E. coli* RNases have been shown to contribute to mRNA degradation (Hajnsdorf *et al.,* 1996), and they are probably all present in the S30 extract. The efficiency of ribosome display was increased approximately 15-fold (Hanes and Plückthun, 1997), when a 5′-stemloop derived from the T7 gene 10 upstream region and a 3′-stemloop derived from the terminator of the *E. coli* lipoprotein were introduced into the ribosome display construct. A similar improvement in efficiency was observed when using the same 5′-stemloop and the 3′-stemloop derived from the early terminator of phage T3 (Hanes and Plückthun, 1997). The stemloop structures may protect the mRNA particularly from degradation by the exonucleases PNPase and RNaseII, which act from the 3′-end of the mRNA (Hajnsdorf *et al.,* 1996), and against RNaseE, which recognizes the 5′-end (Bouvet and Belasco, 1992).

A protein tail, which is the same in all library members, is fused to the C-terminus of the ribosome display construct and serves as a spacer. This spacer has two main functions. First, it tethers the synthesized protein to the ribosome. Second, it keeps the structured part of the protein outside the ribosome and allows its folding and interaction with ligands, without clashing with the ribosomal tunnel. The ribosomal tunnel covers between 20 and 30 C-terminal amino acids of the nascent polypeptide chain during protein synthesis and can therefore prevent the folding of the protein (Malkin and Rich, 1967; Smith *et al.,* 1978).

A ribosome display construct (the library in the ribosome display format) can be prepared completely *in vitro* either by ligation of the DNA library to the spacer region or by assembly PCR of the DNA library and the spacer. All the above-mentioned features, which are important for ribosome display (T7 promoter, ribosome binding site, and stem-loop structure), are then introduced by PCR (Fig. 3).

3. The In Vitro *Translation Step of Ribosome Display*

a. Coupled versus Uncoupled System. The ribosomal complexes can be generated either by a coupled *in vitro* transcription-translation system (Mattheakis *et al.,* 1994; He and Taussig, 1997) using a DNA library as a template, or the mRNA can be first prepared by *in vitro* transcription, purified and subsequently used for *in vitro* translation (Hanes and Plückthun, 1997; Gersuk *et al.,* 1997). The coupled system is much simpler than the uncoupled one, but especially in the case of the *E. coli* S30 system, it yields fewer functional ribosomal complexes (Hanes *et al.,* unpublished experiments). Furthermore, T7 RNA polymerase, which is used for *in vitro* transcription, requires reducing agents such as 2-mercaptoethanol for its stability. If disulfide-containing proteins are displayed in the coupled system, the presence of reducing agents during *in vitro* translation markedly decreases or may even abolish their folding efficiency and thereby the activity of the displayed protein. A separate transcription step under reducing conditions followed by the translation step under oxidizing conditions can be used to solve this dilemma. This problem may in principle also be overcome by preparing T7 RNA polymerase without reducing agent, but enzyme activity has then to be carefully monitored.

b. Time and Temperature of In Vitro *Translation.* Every *in vitro* translation system differs in optimal translation time and temperature, both of which can influence the yield. In the *E. coli in vitro* translation system the translation reaction is usually performed at 37°C (Mattheakis *et al.,* 1994; Hanes and Plückthun, 1997), where the folding efficiency was found to be favorable. Although *in vitro* folding usually gives higher yields at low temperature, the combined temperature effects of the translation reaction, chaperone-assisted folding, escape from nucleases and proteases, and other unknown factors seem to be most favorable at 37°C for the yield of functional proteins.

The time of translation is also very important, especially for uncoupled systems. During *in vitro* translation, protein synthesis follows a saturation curve reaching a plateau after 30 minutes (Ryabova *et al.,* 1997). At the same time the mRNA is continuously degraded with a half-life of approximately 5 to 10 minutes. Thus, an optimal time exists at which the concentration of intact mRNA-ribosome-protein complexes that can be used for selection is maximal. The optimal time for the *E. coli* system is around 7 minutes.

In eukaryotic systems the optimal translation times are usually longer. For instance, an *in vitro* translated truncated lysozyme in a wheat germ system was still present in the ribosomal fraction after 30 minutes of translation (Haeuptle *et al.,* 1986). The translation time in the coupled system is not such a critical parameter, compared to the uncoupled system, since the mRNA is continuously being produced. The reaction time can therefore be extended to 30–60 minutes when using the *E. coli* (Mattheakis *et al.,* 1996) or the rabbit reticulocyte systems (He and Taussig, 1997). A longer translation time is not recommended because some crucial components necessary for translation or transcription become limiting and low molecular weight compounds generated during the translation accumulate and eventually inhibit the *in vitro* translation (Jermutus *et al.,* 1998).

c. Additives to the In Vitro *Translation.* The addition of several components, stabilizing either the mRNA or the ribosomal complexes or improving the protein folding during the *in vitro* translation reaction, can increase the overall efficiency of ribosome display. RNasin, an inhibitor of certain mammalian RNases, was first used in the wheat germ system

(Gersuk *et al.,* 1997), but it was not reported whether it had any effect, and it has no effect in the *E. coli* system.

To stop the translation reaction and further stabilize the ribosomal complexes, cycloheximide can be added in the eukaryotic system (Gersuk *et al.,* 1997). For the same purpose chloramphenicol, an antibiotic that inhibits bacterial protein synthesis by binding to the 23S ribosomal RNA in the peptidyl transferase center, can be used in the *E. coli* system (Mattheakis *et al.,* 1994). However, chloramphenicol was found to have no influence on the efficiency of *E. coli* ribosome display (Hanes and Plückthun, 1997).

Protein disulfide isomerase (PDI) was found to be important in catalyzing disulfide bond formation of antibody fragments, and it improved the efficiency of *E. coli* ribosome display of antibodies threefold when used during the *in vitro* translation reaction (Hanes and Plückthun, 1997) (see also Section II, B). A fourfold improvement of ribosome display was observed when 10Sa-RNA, which is involved in degradation of truncated proteins (see Section III, B, 2), was inhibited by using an antisense DNA oligonucleotide directed against the 10Sa-RNA (Hanes and Plückthun, 1997).

4. *Applications of Ribosome Display*

In contrast to the other *in vitro* selection technologies discussed below, where to date mostly model enrichments have been reported, more examples on directed evolution are available for ribosome display, and these are reported in a separate section (see Section IV). Briefly, experiments on peptide and protein libraries are summarized, and the directed evolution of distinct biophysical properties is discussed.

C. *RNA-Peptide Fusion and* In Vitro *Virus*

1. *Principle of RNA-Peptide Fusion*

A somewhat different approach to couple phenotype and genotype was designed by Roberts and Szostak (1997) and independently by Nemoto *et al.* (1997), who linked a peptide covalently to its encoding mRNA. In this technology, called "RNA-peptide fusion" (Roberts and Szostak, 1997; Roberts, 1999; Roberts and Ja, 1999) or "*in vitro* virus" (Nemoto *et al.,* 1997), mRNA is transcribed *in vitro,* purified, and subsequently ligated at its 3′-terminus to a puromycin-tagged DNA-linker (Fig. 5). This RNA-DNA construct with a puromycin at its 3′-end is again purified and then translated *in vitro.* The ribosome stalls upon reaching the RNA-DNA junction, allowing the puromycin to enter the peptidyltransferase

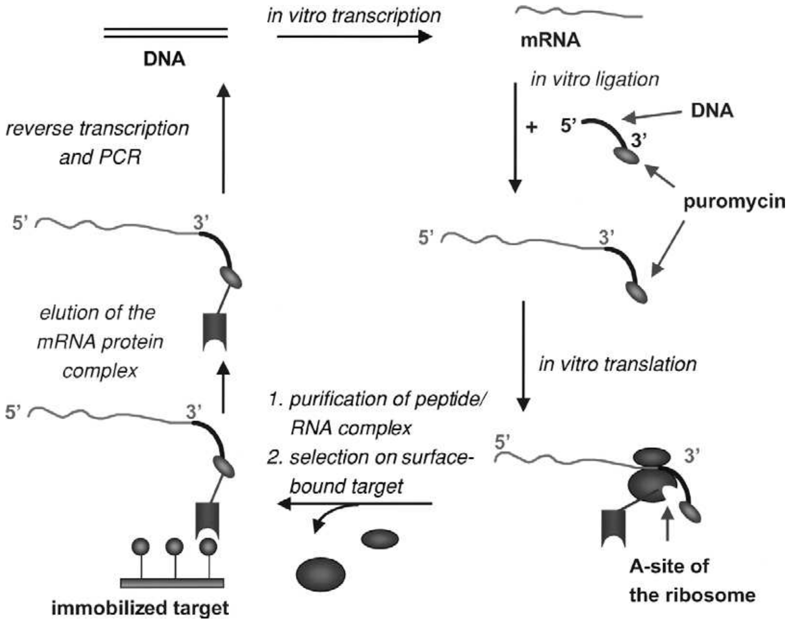


FIG. 5. Protein-RNA fusion. Covalent RNA-protein complexes can be generated by ligation of a DNA-puromycin linker to the *in vitro* transcribed mRNA. During *in vitro* translation, the ribosome stalls at the RNA-DNA junction. Puromycin can then bind to the ribosomal A-site. The nascent polypeptide is thereby transferred to puromycin. The resulting covalently linked complex of mRNA, puromycin, and peptide can be used for selection experiments. After affinity selection, the bound complexes are eluted and subsequently the mRNA is amplified by RT-PCR.

site and covalently couple to the nascent peptide. Optimization of the DNA-linker length minimized cross-contamination, which would result in polypeptides fused to a nonrelated mRNA (Roberts and Szostak, 1997). The resulting covalently linked complex of mRNA-DNA hybrid, puromycin, and the encoded polypeptide can then be dissociated from the ribosome and used for affinity selection. After selection, the bound RNA-peptide fusions can be eluted and amplified by reverse transcription and PCR. One critical and time-consuming step during each cycle is the ligation of the DNA-puromycin linker to the mRNA, which requires careful handling of the mRNA. The advantage of RNA-peptide fusion is that the covalently coupled complexes of mRNA and peptide are robust and therefore should allow more stringent selection conditions than in noncovalently coupled systems, but it is unclear how much the functional library size is decreased by the additional manipulations necessary in this procedure.

2. *Applications of mRNA-peptide Fusion*

In a model enrichment, the DNA encoding a peptide with the myc-epitope was diluted in different ratios in a DNA pool encoding random peptide sequences. The myc-epitope peptide was enriched by immunoprecipitation with an anti-c-myc antibody by a factor of twenty to forty per cycle of mRNA-peptide fusion (Roberts and Szostak, 1997). More recently, the mRNA-peptide fusion approach has been used to select peptides for binding to this anti-c-myc antibody from a library of 2 × 10^{13} molecules (Roberts and Ja, 1999). In these experiments, the enrichment factor was reported to be 200 per round and is therefore similar to the enrichment factors observed with ribosome display.

Protein-ligand interactions can not only be secreened or selected *in vitro,* but also can be directly characterized for particular interaction partners. Nemoto *et al.* (1999) applied the mRNA-peptide fusion technology to fluorescently label the displayed proteins in order to study protein-protein interactions by fluorescence polarization measurements.

D. Selection for Enzymatic Activity In Vitro

A long-standing desire of biochemists has been to generate catalysts, either by design or by selection (Arnold, 1998; Forrer *et al.,* 1999). One strategy, used with some success in the identification of catalytic antibodies, has been to select by phage display molecules that bind to a transition-state analog (Schultz and Lerner, 1995; Arkin and Wells, 1998). Another possibility is to covalently trap phages expressing active catalyst with suicide inhibitors (Baca *et al.,* 1997; Janda *et al.,* 1997).

However, a major limitation of this approach is that appropriate suicide inhibitors or transition state analogs are not available for most enzymatic reactions. Furthermore, binding to the transition-state analog does not necessarily correlate well with catalysis, nor does improved binding automatically generate an improvement in catalysis (Baca *et al.,* 1997).

The fundamental problem with selection for catalysis is that the product of an enzyme leaves the catalytic protein. Thus, even when genetic information of the catalyst is physically connected to the catalytic protein, the phenotype (i.e., the efficiency of the reaction) does not remain connected. In other words, in a mixture of catalysts of different efficiency, the information—which of the molecules has actually produced most of the product—becomes lost. As a consequence, product and enzyme have to remain physically connected for an efficient evolutionary process.

There are two principal ways of achieving this. Either a direct physical link between catalyst and substrate (and thus product) must be created or, alternatively, a compartmentalization of catalyst and replication machinery must be designed. The problem was solved by nature with the emergence of cells as compartments.

In practice, however, it turns out to be difficult to focus selection pressure on just one enzyme in a cellular environment. The challenge of these approaches lies in correctly directing selection pressure (Zhao and Arnold, 1997). Nature is surprisingly inventive in finding other solutions to escape the selection pressure than through the designed library. Many of the cautionary notes about microorganisms evading the selection pressure particularly apply to catalysis.

The selection for catalytic turnover is therefore more complicated than for binding to a transition state analog or a suicide inhibitor. It requires the possibility to physically separate positive library members from the rest (Barbas *et al.,* 1997; Janda *et al.,* 1997; Pedersen *et al.,* 1998). For instance, antibodies against the product can be used to capture product-containing phages (Tawfik *et al.,* 1993). This requires that the substrate and the enzyme are both displayed on the same phage (Pedersen *et al.,* 1998). The substrate can also be linked noncovalently to the phages displaying the protein (Demartis *et al.,* 1999). In principle, the described strategies for selection of catalysts can also be used with *in vitro* techniques such as ribosome display, but this has not yet been reported.

1. Principle of "Cell-like Compartments"

Tawfik and Griffith (1998) reported an *in vitro* selection strategy for catalytic activity using compartmentalization. Here, each member of the DNA library is encapsulated in an aqueous compartment in a water in oil emulsion. The compartments are generated from an *in vitro* transcription-translation system, and contain the components for protein synthesis. The dilution is chosen such that, on average, the water droplets contain less than one DNA molecule. The DNA is transcribed and translated *in vitro* in the presence of substrate, which is covalently attached to the DNA. Only translated proteins with catalytic activity convert the substrate to the product. Subsequently, all DNA molecules are recovered from the water droplets and the DNA linked to the product is separated from the unmodified DNA linked to the educt, which requires a method to discriminate between both. The modified DNA can then be amplified by PCR and used for a second selection cycle. The principle of this approach is depicted in Figure 6.

2. Applications of "Cell-like Compartments"

As a model system, a DNA methylase was chosen (Tawfik and Griffiths, 1998). DNA encoding the methyltransferase activity was methylated by

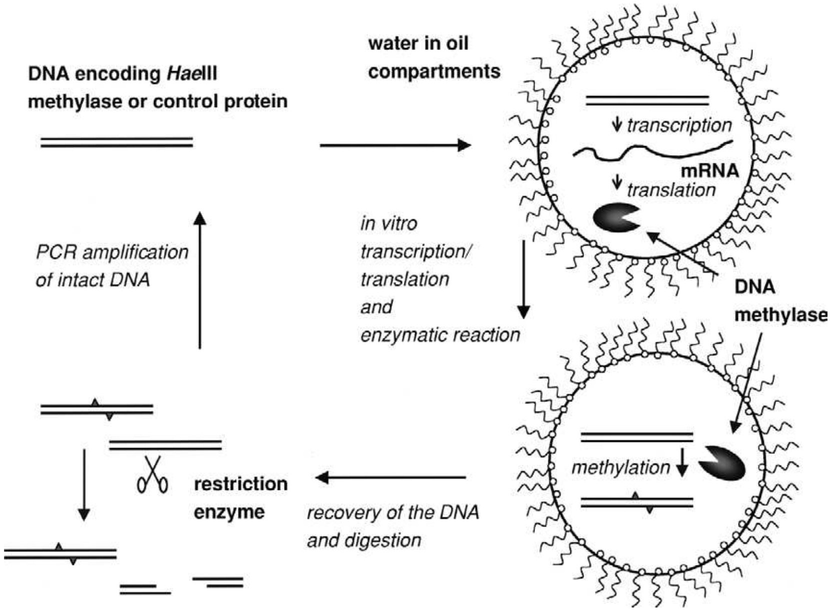

DNA encoding *Hae*III methylase or control protein
water in oil compartments
transcription
mRNA
translation
in vitro transcription/ translation and enzymatic reaction
PCR amplification of intact DNA
DNA methylase
methylation
restriction enzyme
recovery of the DNA and digestion

FIG. 6. Selection for enzymatic activity (DNA methylation) by compartmentalization. DNA encoding *Hae*III methylase is diluted with unrelated DNA (encoding dihydrofolate-reductase). This mixture is dispersed together with a reaction mixture for *in vitro* transcription and translation to form water in oil compartments. The dilution is chosen such that each compartment contains one DNA molecule on average. In the aqueous compartments, the genes are transcribed and translated. In compartments in which an active methylase is translated, the DNA can be methylated and is subsequently recovered from the emulsion and digested by a restriction enzyme. The methylated DNA (encoding *Hae*III methylase) is protected against the digestion, remains intact and is subsequently amplified by PCR.

the enzyme and thereby protected against digestion by a restriction endonuclease. The unmodified DNA was degraded and the intact methylated DNA was amplified by PCR. In a model enrichment, DNA encoding *Hae*III methylase was diluted in different ratios with DNA encoding an irrelevant protein and could be enriched in a single round by a factor of 5000.

The challenge in this technology is to generalize this compartment approach to reactions other than DNA modification. The discrimination of DNA-bound product and DNA linked to educt will be an important step in this endeavor.

E. *Water in Oil Emulsions for Binding Selections (STABLE)*

1. *Principle of Water in Oil Emulsions for Binding Selections (STABLE)*

In an approach similar to the ''cell-like compartments,'' Doi and Yanagawa (1999) used biotinylated DNA to display peptides fused to streptavidin in compartments of water in oil emulsions. The method was named streptavidin-biotin linkage in emulsions, STABLE (Doi and Yanagawa, 1999). Upon *in vitro* translation each translated peptide is displayed as a fusion to streptavidin that binds to its encoding biotinylated DNA in its compartment. The resulting protein-DNA fusions can then be recovered and used for affinity selection. To avoid cross-contamination, biotin has to be added before recovery because much more streptavidin will be produced in each compartment than biotinylated DNA is present. The selected DNA-protein complexes can then be amplified by PCR. The principle of this selection strategy is shown in Figure 7.

One advantage of protein-DNA complexes is that DNA is much more stable as genotype than mRNA. Also, DNA encoding a stop codon can be displayed, facilitating the design of libraries from natural sources (cDNA libraries). However, this advantage may be offset by additional handling steps in this procedure.

It should be noted, however, that streptavidin forms tetramers; therefore four copies of the peptide will be displayed per DNA and a multivalency effect during the selection step is hard to avoid, which may make selection for high affinity more difficult.

2. *Applications of STABLE*

As a model enrichment, Doi and Yanagawa (1999) selected for Ni-NTA binders from a library of decamer peptides. The reported enrichment factor for the his-tag fused to streptavidin was, however, only ten. This is probably due to the low efficiency of the protein DNA fusion formation, which was estimated to be only 1%.

IV. Applications of Ribosome Display

In vitro selection technology can be used in principle for three tasks. We will discuss these in the following sections for ribosome display, as currently the examples are mostly available from this approach. The first task is to identify a protein or a peptide from a pool of variants. Here, ribosome display is used exclusively as a method for *selection,* and changes are not actively introduced in the original pool. In this case, proofreading polymerases are used during the PCR amplification steps (see below). A second application is to *evolve* a given protein or peptide

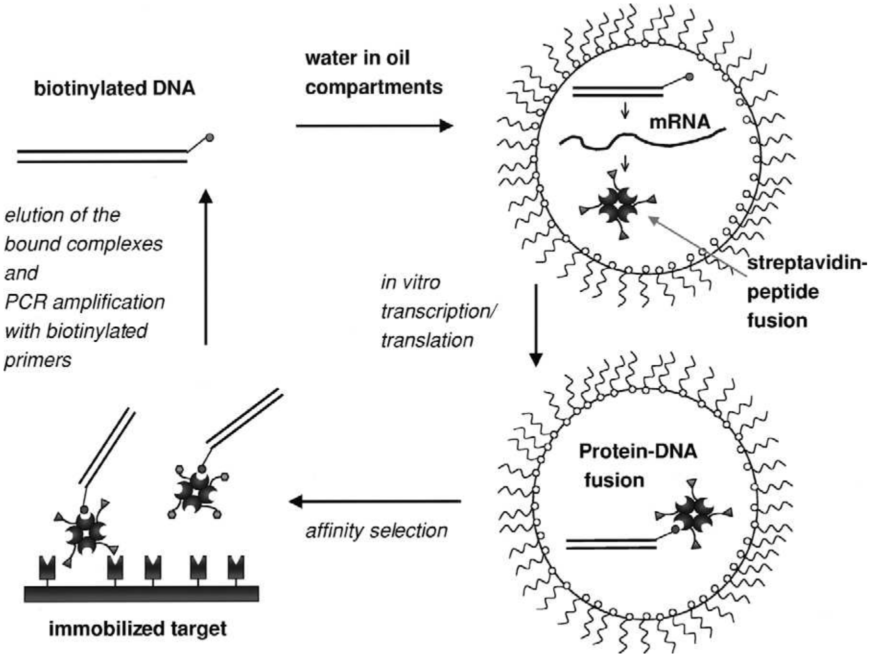


FIG. 7. STABLE. A biotinylated DNA library encoding streptavidin-random peptide fusions is dispersed together with the reaction mixture for *in vitro* transcription and translation in water in oil compartments. The dilution is chosen such that each compartment contains a single DNA molecule on average, which is transcribed and translated. The encoded streptavidin-peptide fusion is synthesized in the compartment and can bind to its encoding biotinylated DNA. The DNA-protein complexes are subsequently recovered from the emulsion and used for affinity selection. The DNA of the bound complexes is then eluted and amplified by PCR.

to obtain new or improved functions or properties. Third, it is possible to combine *both* tasks: select from a pool of variants, when an evolution of the selected clones occurs simultaneously. This requires a diversification of the pool after each selection cycle. In the simplest case, this is accomplished by the inherent error rate of a low-fidelity DNA polymerase used for the PCR amplification steps of ribosome display.

A. *Selection from Libraries with or without Concurrent Sequence Evolution*

1. *Peptide Display*

The first successful application of ribosome display system was demonstrated for the display of a library of random peptide decamers

(Mattheakis *et al.,* 1994). A library of 10^{12} DNA molecules was used with *E. coli* ribosome display utilizing a coupled *in vitro* transcription-translation system. This library was selected for binding to the monoclonal antibody D32.39, which originally bound dynorphin B, a 13-residue opioid peptide, with 0.29 nM affinity. Five cycles of ribosome display resulted in several different peptides with affinities to the antibody ranging from 7.2 to 140 nM affinity. Yet, a peptide with a sequence similar to dynorphin B was not isolated.

A similar approach was followed for displaying a random library of 20-mers using a wheat germ uncoupled transcription and translation system (Gersuk *et al.,* 1997), and several peptides were isolated that bound to prostate-specific antigen, but not to bovine serum albumin. No affinities were reported.

Because the peptide sequences are very short in such systems, few (if any) errors are introduced by the polymerase and even error-prone PCR techniques are not ideal for introducing sequence changes. Therefore, unless a cassette mutagenesis strategy is followed, the benefit of ribosome display for an evolutionary approach is not exploited with displayed peptides. What is exploited, however, is the possibility to use larger libraries than those possible with *in vivo* selection systems.

2. *Protein Display*

The *E. coli* ribosome display system had to be considerably optimized for efficient display of folded proteins. These improvements (explained in more detail in Section III, B) involved the use of RNase inhibitors, the design of hairpins at either end of the RNA and a separate transcription and translation step allowing control over the individual redox requirements, and lead to higher yields, greater stability, and reduced nonspecific binding of the complexes (Hanes and Plückthun, 1997; Hanes *et al.,* 1999).

In a model system of two scFv fragments of an antibody, a 10^9-fold enrichment was achieved by five cycles of ribosome display with an average enrichment of about 100-fold per cycle (Hanes and Plückthun, 1997). All selected scFvs had mutated during five cycles of ribosome display, possessing between zero and four amino acid changes, compared to the original sequence.

Subsequently, it was demonstrated that it is possible to select and evolve scFv antibody fragments from immune libraries using the *E. coli* system. Only three rounds of ribosome display were necessary to isolate a family of scFv fragments binding to a peptide variant of the GCN4 leucine zipper, which exists as a random coil in solution (Hanes *et al.,* 1998). Most of the isolated scFvs had again acquired mutations, with zero to five amino acid changes compared to their consensus sequence, the most likely progenitor scFv, which was present in the library before selection. The best scFv had a dissociation constant of $(4 \pm 1) \times 10^{-11}$ M, measured in solution. The likely common progenitor of the related scFvs bound the antigen with a 65-fold lower affinity than the best binder. This result demonstrated for the first time that from the PCR-based mutants, proteins can be evolved to higher affinity with ribosome display.

The display of proteins using the rabbit reticulocyte ribosome display system was also reported. Here a small library, derived from a scFv derivative of an antibody, V_H-linker-V_L-C_L, binding to progesterone, was used for selection (He and Taussig, 1997). The authors used a coupled *in vitro* transcription-translation system in the presence of 2 mM DTT. However, this concentration of reducing agents can in some cases prevent the folding of disulfide containing proteins such as antibodies.

The eukaryotic ribosome display was subsequently used for selection of human antibody scFv fragments binding progesterone from a library prepared from transgenic mice (He *et al.,* 1999). In this case, a proofreading polymerase was used for the PCR amplification steps included in the ribosome display protocol. It was thereby demonstrated that it is possible to use ribosome display as a method exclusively for selection by virtually maintaining the original library repertoire.

In a direct comparison of the rabbit reticulocyte system to the *E. coli* ribosome display system in a model study, the *E. coli* system turned out to be more efficient for the display of the model scFv constructs tested (Hanes *et al.,* 1999).

A more direct proof that protein variants of higher affinity are truly created during the *E. coli* ribosome display procedure was obtained by using HuCAL (Knappik *et al.,* 2000), a very large synthetic antibody library of 2×10^9 independent members (Hanes *et al.,* 2000). This naive library was applied for six rounds of ribosome display selection using insulin as the antigen. In three independent experiments different scFv families with different framework combinations were isolated. Since the library used was completely synthetic, consisting of forty-nine framework combinations, with the CDR3 regions of both variable domains randomized (Knappik *et al.,* 2000), the starting scFv sequences were known. Thus, any mutations could be directly identified as being generated during the ribosome display procedure. By sequence comparison to the original members of the library it was apparent that all the antibodies selected were not part of the initial library. By using nonproofreading polymerases, mutations were introduced into the enriched pool during the PCR amplification step that is part of each ribosome display cycle.

Thereby, new diversity was generated and each single member of the library began to diversify. This procedure closely mimics the process of somatic hypermutation of antibodies during secondary immunization. The final products of selection are different families of closely related sequences stemming from a common progenitor that started to evolve during ribosome display. A biophysical comparison of the isolated scFvs to their progenitors revealed that all selected scFvs had mutated and significantly improved their affinities to the antigen up to 40-fold by these mutations. The best scFvs had affinities in the low picomolar range (Hanes *et al.,* 2000).

B. *Directed Evolution of Binding Affinity and Stability*

Directed evolution consists of cycles of diversification and selection. Because ribosome display takes place entirely *in vitro,* it can ideally be combined with *in vitro* methods of generating sequence diversity. Since true evolution requires diversification in each cycle, this facile alternation between *in vitro* mutagenesis and *in vitro* selection is one of the attractions of the ribosome display method.

1. *Introducing Diversity*

Depending on the particular protein under consideration, either a focused library or a randomization of the whole gene encoding the protein may be more appropriate. Clearly, this depends on the prior knowledge of the system and its history. If binding affinity is the target function, and the binding site is known, it can be advantageous to first target the residues presumed to be in direct contact with the ligand and then target the whole gene to affect long-range interactions and second sphere residues, which may have subtle effects on the positioning of the direct contact residues. If the binding residues are not exactly known or if they may already be the outcome of an *in vivo* evolution (e.g., in the case of a natural antibody), or in a target function that invariably involves the whole protein—such as stability or folding efficiency—whole gene randomization is probably more effective.

Directed mutagenesis with oligonucleotides encoding mixtures of amino acids (Hermes *et al.,* 1990), error-prone PCR in the presence of nonphysiological metal ions such as Mn^{2+} (Lin-Goerke *et al.,* 1997) or dNTP analogues (Zaccolo *et al.,* 1996) can all be used to randomize the gene either at a particular sequence position or over the whole gene sequence. If a family of homologous proteins with similar function is available, recombination methods such as DNA shuffling (Stemmer, 1994), family shuffling (Crameri *et al.,* 1998), StEP (*St*aggered *e*xtension *p*rocess, Zhao *et al.,* 1998) and RPR (*r*andom-*p*riming *in vitro r*ecombination, Shao *et al.,* 1998) can be used to generate a library. Furthermore, recombination is especially important in between the selection cycles to generate new diversity in the selected pool.

In many of the approaches, the number of mutations per gene can be adjusted, and it is likely that optimal mutation rates exist. On one hand, it is important to have a reasonably large collection of mutants in order to screen sufficient sequence space. On the other hand, it is absolutely necessary to preserve the function of the protein. If too many mutations are introduced, the harmful or destructive mutations can neutralize the beneficial effect of other mutations in one gene. It is generally assumed that *evolution* occurs by steps of increasing fitness and that the sequence of a functional protein must form a continuous path that can be traversed by single mutational steps without passing nonfunctional or less adapted forms (Spetner, 1970; Macken and Pereson, 1989; Smith, 1970; Gillespie, 1984). Favorable double mutations, with one mutation leading to an unfavorable intermediate, have been considered to be rather rare and therefore not important in evolution (Smith, 1970; Gillespie, 1984), even though this is difficult to exclude in general. Therefore, several investigators examined the question of an optimal mutation rate (Kepler and Perelson, 1995 Leigh, 1973; Sasaki and Isawa, 1989) in different contexts. However, as the tolerance of proteins to mutation and the frequency of beneficial mutations is poorly understood at best, further work will be required before this fundamental question can be answered.

2. *Evolving the Affinity of Ligand-binding Proteins*

There are, in principle, two ways to select for improved binding constants. The first strategy relies on equilibrium selection. Here, a displayed library of proteins is incubated with low amounts of ligand such that the concentration of the cognate partner is below the dissociation constant of the protein-ligand complex. On reaching equilibrium, interactions with higher affinity should be favored and, as a consequence, those binders should become enriched during selection. However, the equilibration time τ increases with decreasing ligand concentration:

$$\tau \sim \frac{1}{k_{\mathrm{on}}[L] + k_{\mathrm{off}}} \tag{1}$$

where k_{on} and k_{off} are the kinetic constants and $[L]$ denotes the ligand concentration. If k_{off} is assumed to be very small (e.g., 10^{-6} s^{-1}), which should be the case for tight binders, the first term with ligand concentrations in the picomolar range becomes negligible at a typical k_{on} of 10^5 M^{-1} s^{-1} and the system needs more than a week to equilibrate! Furthermore, weak interactions are not actively excluded from the selection process. These theoretical considerations are in line with experimental observations: In equilibrium selections with the aim of obtaining tighter binders no evolution of the entire sequence pool was observed (Jermutus *et al.,* unpublished observations). Therefore, only very few improved binders could be detected.

Alternatively, the kinetic constants k_{on} and k_{off} can be targeted directly. While k_{on} is primarily controlled by translational and rotational diffusion as well as orientation factors and ranges usually from 10^5 to 10^7 M^{-1} s^{-1} (Northrup and Erickson, 1992), k_{off} of typical receptor-ligand interactions can vary over several order of magnitudes (10^{-1} to 10^{-6} s^{-1}). Off-rate selection has the potential to significantly improve binding affinity (Hawkins *et al.,* 1992; Yang *et al.,* 1995; Boder and Wittrup, 1997, Chen *et al.,* 1999). Furthermore, the selection time can be controlled easily, such that a selection for predefined kinetic constants is feasible. Provided that the ligand can be obtained in sufficient amounts and can be tagged, the protocol for off-rate evolution is straightforward: The displayed protein library is equilibrated with a low concentration of tagged antigen, usually in the range of the starting K_D. In the next step, a high molar excess of free, untagged antigen is added and the incubation is continued for increasing time periods in each evolution round. By adding the free antigen in excess, any dissociation of displayed protein with its tagged ligand becomes irreversible. In this strategy, weakly interacting molecules are titrated from the selection process. Mutants with faster off-rates are actively trapped by free antigen and washed away, even if they are present at very high concentrations. As a consequence, the background of low affinity binding proteins surviving the selection pressure is reduced.

In a directed *in vitro* evolution of antibody affinity (Jermutus *et al.,* 2000), ribosomal complexes coding for the protein library were first equilibrated with nanomolar concentrations of biotinylated antigen, and then an excess of free antigen was added. After time periods that increased in each round, complexes still binding the biotinylated antigen were rescued by the addition of streptavidin magnetic beads. The mRNA coding for these proteins was purified and served as the template for the next evolution round. An initial library of the fluorescein-binding antibody c12 with a starting K_D of 1.2 nM was created by error-prone PCR with the dNTP analogues dPTP {6-(2-deoxy-β-D-ribofuranosyl)-3,4-dihydro-8*H*-pyrimido[4,5-c][1,2]-oxazin-7-one-5′-triphosphate} and 8-oxo-dGTP (Zaccolo *et al.,* 1996). Despite the fact that RNA is generally regarded a labile molecule, the selection could be carried out for more than 10 days, provided that a high magnesium concentration and low temperature were maintained. Three selected mutants were cloned into *E. coli,* expressed and purified.

Measurements of k_{on}, k_{off}, and K_D showed that k_{off} and as a consequence K_D had indeed been improved by more than one order of magnitude (Jermutus *et al.,* 2000). The evolved scFvs all contained multiple mutations, compared to the parent molecule. Between four and eleven amino acids (mean value of 7.2) were mutated per scFv. The majority of these mutations are located in permissive positions on the surface of the molecule, in areas unlikely to be directly involved in antigen recognition. Only two positions are mutated in the majority of all sequences, indicating a strong selection: The mutation of His L94 to tyrosine in CDR L3 affects a residue that points straight into the antigen binding site (Fig. 8A). The mutation of Asp H101 to glycine, alanine, or serine in CDR H3 affects a residue on the outer side of CDR H3, pointing away from the antigen binding pocket. This substitution will have the effect of breaking the salt bridge and increasing the flexibility of CDR H3 such that it can adopt a more ‘‘open’’ conformation (see legend to Fig. 8A for details). The multiple occurrence of Asp H101 replacements shows that this solution for improving the off-rate of the scFv fragment has been found several times independently during the *in vitro* evolution process.

3. Increasing Protein Stability

The energy difference between the native and the denatured state of a protein is very small and typically in the range of 5 to 15 kcal/mol. As a consequence, seemingly minor changes in the structure can lead to reduced stability and/or open up new pathways leading to misfolding. This observation provides the rationale for evolving proteins to higher stabilities *in vitro*: As it is difficult to introduce conditions unfavorable for folding without affecting other components of the *in vitro* system, the best strategy is to reduce the stability of the wild-type protein such that it no longer functions. The selection is then for additional mutations that lead to a regain of function to compensate for the loss incurred by the original destabilization. Mutants with increased folding rates or higher intrinsic stabilities should then be selected. The selection design must be carefully considered, as the level of destabilization will affect the selection background. This background is caused by proteins that survive the selection process without improved properties. The higher the stability of the initial pool, the more difficult it is to select for improved mutants. After a sufficient number of evolution cycles, single mutants that have adapted to the selection stress can be identified. By then removing the stress, such as by reversing a destabilizing mutation or

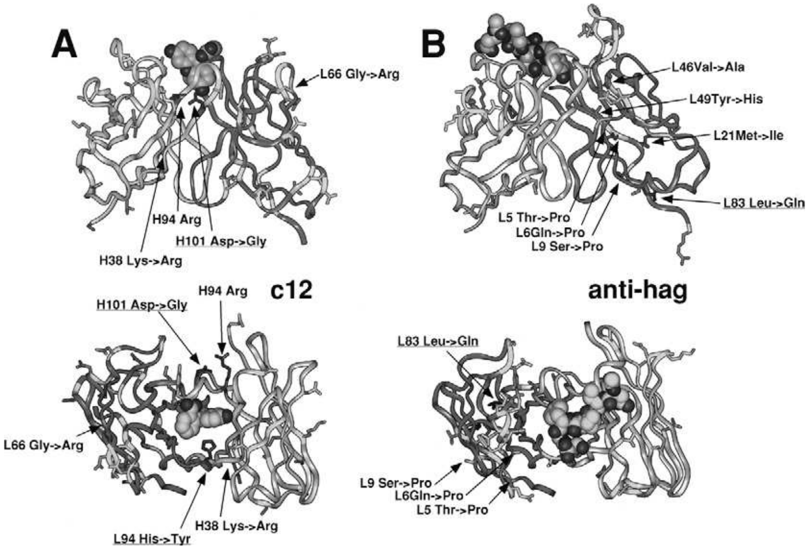


FIG. 8. Localization of the mutated residues after ribosome display. (A) Homology model of the antibody c12 with docked fluorescein. The strongly selected mutations L94 His to Tyr and H101 Asp to Gly, Ser or Ala are labeled and underlined. The mutation in position H101 destroys the salt bridge to Arg H94, which will most likely lead to a more open conformation of CDR3. Other mutations, which have been selected in several clones and are believed to indirectly contribute to binding, are labeled. The substitution of Lys H38 to Arg is very likely to have a stabilizing influence, as it participates in a highly conserved charge cluster with Glu H46 and Asp H86. The mutation of Gly L66 to Arg is expected to have a profound effect on the conformation of the outer loop and thus indirectly affect the geometry of the antigen binding site. (B) Experimental structure of the anti-influenza hemagglutinin antibody Fab 17/9 (PDB entry lifh). The strongly selected mutations L83 Leu to Gln is labeled and underlined. This residue is located at the bottom of the V_L domain and in intact antibodies contributes to the interface between variable and constant domain. However, it is exposed in scFv fragments, and thus a hydrophilic residue here may be beneficial. Mutations in buried and semi-buried positions are concentrated in the V_L domain (Met L21 Ile, Val L46 Ala, Tyr L49 His), while the V_H domain accumulated predominantly surface and interface mutations. In one clone, a stretch of consecutive proline residues (Thr L5 Pro, Gln L6 Pro, Ser L9 Pro, with L8 being a native Pro) was selected. Probably these residues lead to a stabilization by limiting the conformational degrees of freedom of the unfolded state even though they will not be able to form all of the main chain H-bonds of the original sequence and will therefore lead to some conformational adjustments. The positions of the remaining mutated residues are indicated in gray without labeling, illustrating how ribosome display leads to a targeting of the whole sequence.

removing denaturants, more stable variants of the wild-type proteins can be created. To illustrate this method a few examples are given below.

If certain amino acids (such as disulfide bonds or crucial amino acids in the hydrophobic core) are indispensable for protein stability, these positions can be changed by site-directed mutagenesis (Proba *et al.,* 1998). To avoid back-mutations during the evolution process or the selection of a residual wild-type contamination, the pool is amplified after each round of ribosome display with a primer that reintroduces the destabilizing mutation. If the mutation is not close to one of the termini, the coding sequence has to be amplified in two parts, which are then reassembled by PCR. Thus, to evolve improved stabilities this strategy first removes known crucial stabilizing factors to select for compensatory mutations at different positions.

Another approach may be used that focuses on the nascent protein chain. Since ribosome display depends on *in vitro* translation and folding to the functional state, the folding of the nascent chain can, in the case of disulfide-containing proteins, be inhibited by adding DTT (Jermutus *et al.,* 2000). Similar approaches should in principle be possible by adding proteases and suitable amounts of detergents and denaturants, even though this will have to be tested for each particular case. Likewise, important chaperones might be removed from the translation mix by immunoprecipitation. Another, more speculative strategy could make use of published *in vitro* translation systems from extremo- or thermophiles such as *Thermus thermophilus* (Watanabe *et al.,* 1980, Ueda *et al.,* 1991). Here folding, and eventually even selection, could occur at high temperatures.

Single-chain Fv antibody fragments contain two conserved disulfide bridges. These are important stability elements, and the removal of the disulfide bond usually results in a significant loss of activity. Based on this observation a strategy for evolving improved stability was defined (Jermutus *et al.,* 2000). An anti-hemagglutinin scFv with a stability of about 24 kJ/mol and midpoint of denaturation of 4.5 M urea was used as a test case. Because it was shown previously that oxidizing conditions during *in vitro* translation were necessary for maximal yields of functional protein, more stable mutants were selected by choosing a reducing redox potential during the translation step in ribosome display. Over five rounds the selection pressure was gradually increased by increasing the DTT concentration from 0.5 mM to 10 mM, corresponding to a final redox potential of at least −300 mV (assuming less than 0.1% oxidized species at equilibrium). Mutants could only survive the selection process if they folded to a stable structure in the presence of DTT and retained their antigen binding activity. For this purpose, ribosomal complexes were incubated under reducing conditions on immobilized antigen (hag-peptide) and washed only briefly to avoid any selection for tighter binders. Since the selection is designed to enrich mutants that regain sufficient functionality, higher affinities are a possible selection shortcut (Jung *et al.,* 1999). After five selection rounds single mutants were cloned into *E. coli,* expressed, and purified. The most stable protein had increased its stability by about 30 kJ/mol, shifted its denaturation midpoint by 0.9 M urea, and displayed an m-value from the denaturation curve very close to the theoretical value for a two-state transition. From urea renaturation experiments under reducing conditions it was concluded that the evolved mutants could quantitatively refold in the presence of DTT. In contrast, the transition of the wild-type protein under reducing conditions indicated a population of nonnative species remaining after refolding and, thus, incomplete reversibility. Moreover, unlike the wild-type protein, all the mutants could be functionally expressed in the cytoplasm. Sequencing revealed that the mutants had acquired three to seven mutations in the coding sequence (mean value of 4.8). From modeling (Fig. 8B) and biophysical analysis of these mutants it could be concluded that they had all used different mutation strategies to adapt to the selection pressure. Both experiments, the maturation of off-rate and stability, indicated that the mutants had used different lineages during the evolution process, which is most probably due to the large library size in each ribosome display selection.

V. Perspectives of Directed *In Vitro* Evolution

The *in vitro* evolution of proteins is now a reality. To date, most of the evolutionary experiments have been carried out with ribosome display, but applications of the other described technologies will surely follow. The increased library size during selection and the experimental ease of including complex diversification techniques make *in vitro* selection techniques the methods of choice for the deliberate alteration of protein characteristics. However, *in vitro* selections will only be successful if protocols can be designed that direct the evolution process to the intended phenotype and minimize the risk of selection shortcuts. While selection for binding, improved affinity, and increased stability have now been described in the literature, more challenging goals such as enzymatic activity, expressability, or, in the case of scFvs, shifts in monomer-dimer equilibrium will need considerable effort for designing generally applicable selection strategies.

It is likely that future developments will address the automation of *in vitro* evolution technologies, as chemical processes are in general easier to automate than biological ones. Cellular processes are variable and dependent on more parameters, some of which are difficult or impossible to influence.

In vitro protein evolution will complement but not replace hypothesis-driven or structure-aided engineering, as it would be very uneconomical to not make use of available knowledge regarding crucial residues, interactions, or known structural transitions. Directed evolution is suited to making the small adjustments that are beyond today's predictive methods. Further, instead of repeatedly discovering the same features in every directed evolution experiment, it can be advantageous to ''dope'' a library with possible mutations in the suspected positions. Thus, a combination of both methods—structure-based rough sketching and evolutionary fine-tuning—is likely to become a standard approach for solving practical problems in protein engineering and design. A very important corollary of this perspective is that there is great merit in detailed biophysical study of the effect of point mutations, as this knowledge will greatly facilitate the design of ''smart libraries.''

In conclusion, there are four key advantages in carrying out selections and evolutionary refinements *in vitro.* First, it is rapid, as no cellular cloning is involved. Second, the size of libraries is only limited by the amount of DNA (or RNA) that can be handled. Third, it is, in general, easier to focus the selection pressure on the quantity in question *in vitro* than in the highly variable context of a living cell. Fourth, the interfacing of selection from complex libraries and their simultaneous evolution is more convenient, as both can be carried out *in vitro. In vitro* protein evolution thus has a bright future.

Acknowledgments

C. S. and L. J. are supported by a predoctoral Kékule-fellowship from the Fonds der Chemischen Industrie (Germany). The authors would like to thank Stephen Marino, Julia Henckel, and Imre Berger for critical review of the manuscript.

References

Allen, D., Simon, T., Sablitzky, F., Rajewsky, K., and Cumano, A. (1988). *EMBO J.* **7,** 1995–2001.

Andersen, P. S., Stryhn, A., Hansen, B. E., Fugger, L., Engberg, J., and Buus, S. (1996). *Proc. Natl. Acad. Sci. USA* **93,** 1820–1824.

Arkin, M. R., and Wells, J. A. (1998). *J. Mol. Biol.* **284,** 1083–1094.

Arnold, F. H. (1998). *Nat. Biotechnol.* **16,** 617–618.

Baca, M., Scanlan, T. S., Stephenson, R. C., and Wells, J. A. (1997). *Proc. Natl. Acad. Sci. USA* **94,** 10063–10068.

Barbas, C. F., Heine, A., Zhong, G., Hoffmann, T., Gramatikova, S., Bjornestedt, R., List, B., Anderson, J., Stura, E. A., Wilson, I. A., and Lerner, R. A. (1997). *Science* **278,** 2085–2092.

Barbas, C. F., Bain, J. D., Hoekstra, D. M., and Lerner, R. A. (1992). *Proc. Natl. Acad. Sci. USA* **89,** 4457–4461.

Berek, C., and Milstein, C. (1987). *Immunol. Rev.* **96,** 23–41.

Berger, S. L., and Birkenmeier, C. S. (1979). *Biochemistry* **18,** 5143–5149.

Boder, E. T., and Wittrup, K. D. (1997). *Nat. Biotechnol.* **15,** 553–557.

Bouvet, P., and Belasco, J. G. (1992). *Nature* **360,** 488–491.

Cadwell, R. C., and Joyce, G. F. (1992). *PCR Methods Appl.* **2,** 28–33.

Chen, Y., Wiesmann, C., Fuh, G., Li, B., Christinger, H. W., McKay, P., de Vos, A. M., and Lowman, H. B. (1999). *J. Mol. Biol.* **293**, 865–881.

Chien, C. T., Bartel, P. L., Sternglanz, R., and Fields, S. (1991). *Proc. Natl. Acad. Sci. USA* **88,** 9578–9582.

Chothia, C., and Lesk, A. M. (1987). *J. Mol. Biol.* **196,** 901–917.

Crameri, A., Raillard, S. A., Bermudez, E., and Stemmer, W. P. (1998). *Nature* **391,** 288–291.

Daugherty, P. S., Olsen, M. J., Iverson, B. L., and Georgiou, G. (1999) *Protein Eng.* **12,** 613–621.

Demartis, S., Huber, A., Viti, F., Lozzi, L., Giovannoni, L., Neri, P., Winter, G., and Neri, D. (1999). *J. Mol. Biol.* **286,** 617–633.

De Pasquale, C., and Kanduc, D. (1998). *Biochem. Mol. Biol. Int.* **45,** 1005–1009.

Doi, N., and Yanagawa, H. (1999). *FEBS Lett.* **457,** 227–230.

Dougan, D. A., Malby, R. L., Gruen, L. C., Kortt, A. A., and Hudson, P. J. (1998). *Protein Eng.* **11,** 65–74.

Dower, W. J., and Cwirla, S. E. (1992). In: Chang, D. C., Chassy, B. M., Saunders, J. A., and Sowers, A. E. (eds.), *Guide to electroporation and electrofusion.* Academic Press, San Diego, 291.

Eaton, B. E., Gold, L., Hicke, B. J., Janjic, N., Jucker, F. M., Sebesta, D. P., Tarasow, T. M., Willis, M. C., and Zichi, D. A. (1997). *Bioorg. Med. Chem.* **5,** 1087–1096.

Eigen, M., and Rigler, R. (1994). *Proc. Natl. Acad. Sci. USA* **91,** 5740–5747.

Engel, A., Lyubchenko, Y., and Müller, D. (1999). *Trends Cell Biol.* **9,** 77–80.

Fields, S., and Song, O. (1989). *Nature* **340,** 245–246.

Forrer, P., Jung, S., and Plückthun, A. (1999) *Curr. Opin. Struct. Biol.* **9,** 514–520.

Frolova, L., Le Goff, X., Rasmussen, H. H., Cheperegin, S., Drugeon, G., Kress, M., Arman, I., Haenni, A. L., Celis, J. E., Philippe, M., *et al.* (1994). *Nature* **372,** 701–703.

Georgiou, G., Poetschke, H. L., Stathopoulos, C., and Francisco, J. A. (1993). *Trends Biotechnol.* **11,** 6–10.

Gersuk, G. M., Corey, M. J., Corey, E., Stray, J. E., Kawasaki, G. H., and Vessella, R. L. (1997). *Biochem. Biophys. Res. Commun.* **232,** 578–582.

Gillespie, J. H. (1984). *Evolution* **38,** 1116–1129.

Gold, L., Polisky, B., Uhlenbeck, O., and Yarus, M. (1995). *Ann. Rev. Biochem.* **64,** 763–797.

Gram, H., Marconi, L. A., Barbas, C. F., Collet, T. A., Lerner, R. A., and Kang, A. S. (1992). *Proc. Natl. Acad. Sci. USA* **89,** 3576–3580.

Grentzmann, G., Brechemier-Baey, D., Heurgue-Hamard, V., and Buckingham, R. H. (1995). *J. Biol. Chem.* **270,** 10595–10600.

Griffiths, A. D., Williams, S. C., Hartley, O., Tomlinson, I. M., Waterhouse, P., Crosby, W. L., Kontermann, R. E., Jones, P. T., Low, N. M., Allison, T. J., *et al.* (1994). *EMBO J.* **13,** 3245–3260.

Haeuptle, M. T., Frank, R., and Dobberstein, B. (1986). *Nucleic Acids Res.* **14,** 1427–1448.

Hajnsdorf, E., Braun, F., Haugel-Nielsen, J., Le Derout, J., and Regnier, P. (1996). *Biochimie* **78,** 416–424.

Hanes, J., Schaffitzel, C., Knappik, A., and Plückthun, A. (2000). Submitted.

Hanes, J., Jermutus, L., Schaffitzel, C., and Plückthun, A. (1999). *FEBS Lett.* **450,** 105–110.

Hanes, J., Jermutus, L., Weber-Bornhauser, S., Bosshard, H. R., and Plückthun, A. (1998). *Proc. Natl. Acad. Sci. USA* **95,** 14130–14135.

Hanes, J., and Plückthun, A. (1997). *Proc. Natl. Acad. Sci. USA* **94,** 4937–4942.

Hawkins, R. E., Russell, S. J., and Winter, G. (1992). *J. Mol. Biol.* **226,** 889–896.

He, M., Menges, M., Groves, M. A., Corps, E., Liu, H., Brüggemann, M., and Taussig, M. J. (1999). *J. Immunol. Methods* **231,** 105–117.

He, M., and Taussig, M. J. (1997). *Nucleic Acids Res.* **25,** 5132–5134.

Hellinga, H. W. (1997). *Proc. Natl. Acad. Sci. USA* **94,** 10015–10017.

Hentze, M. W. (1995). *Curr. Opin. Cell Biol.* **7,** 393–398.

Hermann, T., and Patel, D. J. (2000). *Science* **287,** 820–825.

Hermes, J. D., Blacklow, S. C., and Knowles, J. R. (1990). *Proc. Natl. Acad. Sci. USA* **87,** 696–700.

Hoogenboom, H. R. (1997). *Trends Biotechnol.* **15,** 62–70.

Hoogenboom, H. R., and Winter, G. (1992). *J. Mol. Biol.* **227,** 381–388.

Hurle, M. R., and Gross, M. (1994). *Curr. Opin. Biotechnol.* **5,** 428–433.

Jakobovits, A., Green, L. L., Hardy, M. C., Maynard-Currie, C. E., Tsuda, H., Louie, D. M., Mendez, M. J., Abderrahim, H., Noguchi, M., Smith, D. H., *et al.* (1995). *Ann. N. Y. Acad. Sci.* **764,** 525–535.

Janda, K. D., Lo, L. C., Lo, C. H. L., Sim, M. M., Wang, R., Wong, C. H., and Lerner, R. A. (1997). *Science* **275,** 945–948.

Janosi, L., Shimizu, I., and Kaji, A. (1994). *Proc. Natl. Acad. Sci. USA* **91,** 4249–4253.

Jermutus, L., Honegger, A., Schwesinger, F., Hanes, J., and Plückthun, A. (2000). Submitted.

Jermutus, L., Ryabova, L. A., and Plückthun, A. (1998). *Curr. Opin. Biotechnol.* **9,** 534–548.

Jung, S., Honegger, A., and Plückthun, A. (1999). *J. Mol. Biol.* **294,** 163–180.

Kanduc, D. (1997). *Arch. Biochem. Biophys.* **342,** 1–5.

Kasahara, N., Dozy, A. M., and Kan, Y. W. (1994). *Science* **266,** 1373–1376.

Kawasaki, G. (1991). *PCT Int. Appl.* WO 91/05058.

Keiler, K. C., Waller, P. R., and Sauer, R. T. (1996). *Science* **271,** 990–993.

Kepler, T. B., and Perelson, A. S. (1995). *Proc. Natl. Acad. Sci. USA* **92,** 8219–8223.

Kieke, M. C., Cho, B. K., Boder, E. T., Kranz, D. M., and Wittrup, K. D. (1997). *Protein Eng.* **10,** 1303–1310.

Kim, D. M., and Schwartz, J. R. (1999). *Biotechnol. Bioeng.* **66,** 180–188.

Knappik, A., Ge, L., Honegger, A., Pack, P., Fischer, M., Wellnhofer, G., Hoess, A., Wolle, J., Plückthun, A., and Virnekäs, B. (2000). *J. Mol. Biol.* **296,** 57–86.

Köhler, G., and Milstein, C. (1975). *Nature* **256,** 495–497.

Komar, A. A., Lesnik, T., and Reiss, C. (1999). *FEBS Lett* **462,** 387–391.

Komine, Y., Kitabatake, M., Yokogawa, T., Nishikawa, K., and Inokuchi, H. (1994). *Proc. Natl. Acad. Sci. USA* **91,** 9223–9227.

Korman, A. J., Knudsen, P. J., Kaufman, J. F., and Strominger, J. L. (1982). *Proc. Natl. Acad. Sci. USA* **79,** 1844–1848.

Kozak, M. (1989). *Mol. Cell. Biol.* **9,** 5134–5142.

Kozak, M. (1984). *Nature* **308,** 241–246.

Kraus, J. P., and Rosenberg, L. E. (1982). *Proc. Natl. Acad. Sci. USA* **79,** 4015–4019.

Kristensen, P., and Winter, G. (1998). *Fold. Des.* **3,** 321–328.

Kuchner, O., and Arnold, F. H. (1997). *Tibtech* **15,** 523–530.

Lancet, D., Sadovsky, E., and Seidemann, E. (1993). *Proc. Natl. Acad. Sci. USA* **90,** 3715–3719.

Leigh, E. G., Jr. (1973). *Genetics* **73,** Suppl 73:1–18.

Lin-Goerke, J. L., Robbins, D. J., and Burczak, J. D. (1997). *Biotechniques* **23,** 409–412.

Lonberg, N., Taylor, L. D., Harding, F. A., Trounstine, M., Higgins, K. M., Schramm, S. R., Kuo, C. C., Mashayekh, R., Wymore, K., McCabe, J. G., *et al.* (1994). *Nature* **368,** 856–859.

Low, N. M., Holliger, P. H., and Winter, G. (1996). *J. Mol. Biol.* **260,** 359–568.

Macken, C. A., and Perelson, A. S. (1989). *Proc. Natl. Acad. Sci. USA* **86,** 6191–6195.

Makeyev, E. V., Kolb, V. A., and Spirin, A. S. (1999). *FEBS Lett.* **444,** 177–180.

Malkin, L. I., and Rich, A. (1967). *J. Mol. Biol.* **26,** 329–346.

Marks, J. D., Hoogenboom, H. R., Bonnert, T. P., McCafferty, J., Griffiths, A. D., and Winter, G. (1991). *J. Mol. Biol.* **222,** 581–597.

Mattheakis, L. C., Bhatt, R. R., and Dower, W. J. (1994). *Proc. Natl. Acad. Sci. USA* **91,** 9022–9026.

Mattheakis, L. C., Dias, J. M., and Dower, W. J. (1996). *Methods Enzymol.* **267,** 195–207.

Matthews, D. J., and Wells, J. A. (1993). *Science* **260,** 1113–1117.

Moore, J. C., Jin, H. M., Kuchner, O., and Arnold, F. H. (1997). *J. Mol. Biol.* **272,** 336–347.

Nemoto, N., Miyamoto-Sato, E., Husimi, Y., and Yanagawa, H. (1997). *FEBS Lett.* **414,** 405–408.

Nemoto, N., Miyamoto-Sato, E., and Yanagawa, H. (1999). *FEBS Lett.* **462,** 43–46.

Nissim, A., Hoogenboom, H. R., Tomlinson, I. M., Flynn, G., Midgley, C., Lane, D., and Winter, G. (1994). *EMBO J.* **13,** 692–698.

Northrup, S. H., and Erickson, H. P. (1992). *Proc. Natl. Acad. Sci. USA* **89,** 3338–3342.

Osborne, S. E., Matsumura, I., and Ellington, A. D. (1997). *Curr. Opin. Chem. Biol.* **1,** 5–9.

Pedersen, H., Holder, S., Sutherlin, D. P., Schwitter, U., King, D. S., and Schultz, P. G. (1998). *Proc. Natl. Acad. Sci. USA* **95,** 10523–10528.

Pelletier, J. N., Arndt, K. M., Plückthun, A., and Michnick, S. W. (1999). *Nat. Biotechnol.* **17,** 683–690.

Pelletier, J. N., Campbell-Valois, F. X., and Michnick, S. W. (1998). *Proc. Natl. Acad. Sci. USA* **95,** 12141–12146.

Proba, K., Wörn, A., Honegger, A., and Plückthun, A. (1998). *J. Mol. Biol.* **275,** 245–253.

Ray, B. K., and Apirion, D. (1979). *Mol. Gen. Genet.* **174,** 25–32.

Rigler, R. (1995). *J. Biotechnol.* **41,** 177–186.

Roberts, R. W. (1999). *Curr. Opin. Chem. Biol.* **3,** 268–273.

Roberts, R. W., and Ja, W. W. (1999). *Curr. Opin. Struct. Biol.* **9,** 521–529.

Roberts, R. W., and Szostak, J. W. (1997). *Proc. Natl. Acad. Sci. USA* **94,** 12297–12302.

Roche, E. D., and Sauer, R. T. (1999). *EMBO J.* **18,** 4579–4589.

Roth, A., and Breaker, R. R. (1998). *Proc. Natl. Acad. Sci. USA* **95,** 6027–6031.

Ruan, B., Hoskins, J., Wang, L., and Bryan, P. N. (1998). *Protein Sci.* **7,** 2345–2353.

Rubingh, D. N. (1997). *Curr. Opin. Biotechnol.* **8,** 417–422.

Ruckman, J., Green, L. S., Beeson, J., Waugh, S., Gillette, W. L., Henninger, D. D., Claesson-Welsh, L., and Janjic, N. (1998). *J. Biol. Chem.* **273,** 20556–20567.

Ryabova, L. A., Desplancq, D., Spirin, A. S., and Plückthun, A. (1997). *Nat. Biotechnol.* **15,** 79–84.

Sasaki, A., and Isawa, Y. (1989). *Theor. Popul. Biol.* **39,** 201–239.

Schier, R., and Marks, J. D. (1996). *Hum. Antibodies Hybridomas* **7,** 97–105.

Schmidt-Dannert, C., and Arnold, F. H. (1999). *Trends Biotechnol.* **17,** 135–136.

Schultz, P. G., and Lerner, R. A. (1995). *Science* **269,** 1835–1842.

Shao, Z., Zhao, H., Giver, L., and Arnold, F. H. (1998). *Nucleic Acids Res.* **26,** 681–683.

Shen, X.-C., Yao, S.-L., Terada, S., Nagamune, T., and Suzuki, E. (1998). *Biochem. Eng. J.* **2,** 23–28.

Shine, J., and Dalgarno, L. (1975). *Nature* **254,** 34–38.

Sieber, V., Plückthun, A., and Schmid, F. X. (1998). *Nat. Biotechnol.* **16,** 955–960.

Smith, G. P. (1985). *Science* **228,** 1315–1317.

Smith, W. P., Tai, P. C., and Davis, B. D. (1978). *Proc. Natl. Acad. Sci. USA* **75,** 5922–5925.

Spada, S., Honegger, A., and Plückthun, A. (1998). *J. Mol. Biol.* **283,** 395–407.

Spetner, L. M. (1970). *Nature* **226,** 948–949.

Stemmer, W. P. (1994). *Nature* **370,** 389–391.

Tarasow, T. M., Tarasow, S. L., and Eaton, B. E. (1997). *Nature* **389,** 54–57.

Tate, W. P., and Brown, C. M. (1992). *Biochemistry* **31,** 2443–2450.

Tawfik, D. S., Green, B. S., Chap, R., Sela, M., and Eshhar, Z. (1993). *Proc. Natl. Acad. Sci. USA* **90,** 373–377.

Tawfik, D. S., and Griffiths, A. D. (1998). *Nat. Biotechnol.* **16,** 652–656.

Thanaraj, T. A., and Argos, P. (1996). *Protein Sci.* **5,** 1594–1612.

Tuerk, C., and Gold, L. (1990). *Science* **249,** 505–510.

Tuite, M. F., and Stansfield, I. (1994). *Mol. Biol. Rep.* **19,** 171–181.

Ueda, T., Tohda, H., Chikazumi, N., Eckstein, F., and Watanabe, K. (1991). *Nucleic Acids Res.* **19,** 547–552.

Vaughan, T. J., Williams, A. J., Pritchard, K., Osbourn, J. K., Pope, A. R., Earnshaw, J. C., McCafferty, J., Hodits, R. A., Wilton, J., and Johnson, K. S. (1996). *Nat. Biotechnol.* **14,** 309–314.

Watanabe, K., Oshima, T., Iijima, K., Yamaizumi, Z., and Nishimura, S. (1980). *J. Biochem.* (*Tokyo*) **87**(1), 1–13.

Winter, G., Griffiths, A. D., Hawkins, R. E., and Hoogenboom, H. R. (1994). *Ann. Rev. Immunol.* **12,** 433–455.

Yang, W. P., Green, K., Pinz-Sweeney, S., Briones, A. T., Burton, D. R., and Barbas, C. F., 3rd (1995). *J. Mol. Biol.* **254,** 392–403.

Yelton, D. E., Rosok, M. J., Cruz, G., Cosand, W. L., Bajorath, J., Hellstrom, I., Hellstrom, K. E., Huse, W. D., and Glaser, S. M. (1995). *J. Immunol.* **155,** 1994–2004.

Yu, A., Barreiro, V., and Haggard-Ljungquist, E. (1994). *J. Virol.* **68,** 4220–4226.

Zaccolo, M., Williams, D. M., Brown, D. M., and Gherardi, E. (1996). *J. Mol. Biol.* **255,** 589–603.

Zhao, H., and Arnold, F. H. (1997). *Curr. Opin. Struct. Biol.* **7,** 480–485.

Zhao, H., Giver, L., Shao, Z., Affholter, J. A., and Arnold, F. H. (1998). *Nat. Biotechnol.* **16,** 258–261.

Zhouravleva, G., Frolova, L., Le Goff, X., Le Guellec, R., Inge-Vechtomov, S., Kisselev, L., and Philippe, M. (1995). *EMBO J.* **14,** 4065–4072.

This Page Intentionally Left Blank

AUTHOR INDEX

SUBJECT INDEX